View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Research Papers in Economics

NBER WORKING PAPER SERIES

OVER-THE-COUNTER MARKETS

Darrell Duffie
Nicolae Garleanu
Lasse Heje Pedersen

Working Paper 10816
http://www.nber.org/papers/w10816

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2004

Part of this paper was previously distributed under the title “Valuation in Dynamic Bargaining Markets.” We are grateful for conversations with Yakov Amihud, Helmut Bester, Joseph Langsam of Morgan Stanley DeanWitter, Richard Lyons, Tano Santos, and Jeff Zwiebel, and to participants at the NBER Asset Pricing Meeting, the Cowles Foundation Incomplete Markets and Strategic Games Conference, theWestern Finance Association conference, the CEPR meeting at Gerzensee, University College London, Universite Libre de Bruxelles, Tel Aviv University, and Universitat Autonoma de Barcelona. The views expressed herein are those of the author(s) and not necessarily those of the National Bureau of Economic Research.

©2004 by Darrell Duffie, Nicolae Garleanu, and Lasse Heje Pedersen. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Over-the-Counter Markets
Darrell Duffie, Nicolae Garleanu, and Lasse Heje Pedersen
NBER Working Paper No. 10816
September 2004
JEL No. G0, G1

ABSTRACT

We study how intermediation and asset prices in over-the-counter markets are affected by illiquidity associated with search and bargaining. We compute explicitly the prices at which investors trade with each other as well as marketmakers' bid and ask prices in a dynamic model with strategic agents. Bid-ask spreads are lower if investors can more easily find other investors, or have easier access to multiple marketmakers. With a monopolistic marketmaker, bid-ask spreads are higher if investors have easier access to the marketmaker. We characterize endogenous search and welfare, and discuss empirical implications.

Darrell Duffie
Graduate School of Business
Stanford University
Stanford, CA 94305-5015
and NBER
duffie@stanford.edu

Nicolae Garleanu
Wharton School
University of Pennsylvania
3620 Locust Walk
Philadelphia, PA 19104-6367
garleanu@wharton.upenn.edu

Lasse Heje Pedersen
NYU Stern Finance
44 West Fouth Street
Suite 9-190
New York, NY 10012
and NBER
lpederse@stern.nyu.edu

In over-the-counter (OTC) markets, an investor who wishes to sell must search for a buyer, incurring opportunity or other costs until one is found. Some OTC markets therefore have intermediaries. Contact with relevant intermediaries, however, is not immediate. Often, intermediaries must be approached sequentially. Hence, when two counterparties meet, their bilateral relationship is inherently strategic. Prices are set through a bargaining process that reflects each investor's or marketmaker's alternatives to immediate trade.

These search-and-bargaining features are empirically relevant in many markets, such as those for mortgage-backed securities, corporate bonds, emerging-market debt, bank loans, derivatives, and certain equity markets. Also, real-estate values are influenced by imperfect search, the relative impatience of investors for liquidity, outside options for trade, and the role and profitability of brokers.

We build a dynamic asset-pricing model capturing these features, and analytically derive the equilibrium allocations, prices negotiated between investors, as well as marketmakers' bid and ask prices. We show how these equilibrium properties depend on investors' search abilities, marketmaker accessibility, and bargaining powers. We determine the search intensities that marketmakers choose, and derive the associated welfare implications of investment in marketmaking.

Our model of search is a variant of the coconuts model of Diamond (1982).[1] A continuum of investors contact each other, independently, at some mean rate λ, a parameter reflecting search ability. Similarly, marketmakers contact agents at some intensity ρ that reflects dealer availability. When agents meet, they bar-

[1]Our model differs from Diamond (1982), and the labor literature more generally, by considering repeated trade of long-lived assets. The monetary search literature (for example, Kiyotaki and Wright (1993)) also considers long-lived assets, but, with the exception of Trejos and Wright (1995), it considers exogenous prices. Our model has similarities with that of Trejos and Wright (1995), but their objectives are different and they do not study marketmaking. See also Harris (1979).

gain over the terms of trade. Gains from trade arise from heterogeneous costs or benefits of holding assets. For example, an asset owner can be anxious to sell because of a liquidity need or because of hedging motives. Marketmakers can offload their inventories in a frictionless inter-dealer market and trade with investors, capturing part of the difference between the inter-dealer price and investors' reservation values.

Asset pricing with exogenous trading frictions has been studied by Amihud and Mendelson (1986), Constantinides (1986), and Vayanos (1998). We endogenize the trading frictions arising through search and bargaining, and show their effects on asset prices. In follow-up work, Duffie, Gârleanu, and Pedersen (2003) extend the model developed here in order to characterize the impact on asset pricing of search in settings with risk aversion and risk limits, while Weill (2002) and Vayanos and Wang (2002) consider cross-sectional asset pricing in extensions with multiple assets.

Market frictions have been used to explain the existence and behavior of marketmakers. Notably, marketmakers' bid and ask prices have been explained by inventory considerations (Garman (1976), Amihud and Mendelson (1980), and Ho and Stoll (1981)), and by adverse selection arising from asymmetric information (Bagehot (1971), Glosten and Milgrom (1985), and Kyle (1985)). In contrast, we model marketmakers who have no inventory risk because of the existence of inter-dealer markets, and our agents are symmetrically informed. In our model, bid and ask prices are set in light of investors' outside options, which reflect both the accessibility of other marketmakers and investors' own abilities to find counterparties.

We show that bid-ask spreads are lower if investors can find each other more

easily.[2] The intuition is that improving an investor's search alternatives forces marketmakers to give better prices. This result is supported by the experimental evidence of Lamoureux and Schnitzlein (1997).

An investor also improves his bargaining position relative to a marketmaker if he can more easily find other marketmakers. Hence, despite the bilateral nature of bargaining between a marketmaker and an investor, marketmakers are effectively in competition with each other over order flow, given the option of investors to search for better terms. Consistent with this intuition, we prove that competitive prices and vanishing spreads obtain as marketmakers' contact intensities become large, provided that marketmakers do not have all of the bargaining power.

In summary, if investors are more sophisticated (that is, have better access to other investors or to marketmakers who do not have total bargaining power), they receive a *tighter* bid-ask spread. This implication sets our theory of intermediation apart from information-based models, in which more sophisticated (that is, better informed) investors receive a wider bid-ask spread.

In an extension with heterogeneous investors in the same OTC market, we show that more sophisticated investors (those with better access to marketmakers) receive tighter bid-ask spreads because of their improved outside options. Hence, this result holds both when comparing across markets and when comparing across investors in the same market. This sets our theory apart from inventory-based models, since these would not imply differential treatment across investors.[3] Fur-

[2]We show that our model specializes in a specific way to the standard general-equilibrium paradigm as bilateral trade becomes increasingly active, under conditions to be described, extending a chain of results by Rubinstein and Wolinsky (1985), Gale (1987), Gale (1986a), Gale (1986b), and McLennan and Sonnenschein (1991), in a manner explained later in our paper. Thus, "standard" asset-pricing theory is not excluded, but rather is found at the end of the spectrum of increasingly "active" markets.

[3]We note that, when comparing across markets, inventory considerations may have the same bid-ask implication as our search model, because more frequent meetings between investors and

ther, in the heterogeneous-agents extension, investors with lower search ability may refrain entirely from trade.

Our result seems consistent with behavior in certain OTC markets, such as those for interest-rate swaps and foreign exchange, in which asymmetric information is limited. Anecdotal evidence suggests that "sales traders" give more competitive prices to sophisticated investors, perceived to have better outside options.

We also consider the case in which the marketmaker has total bargaining power. The bid-ask spread of such a monopolistic marketmaker vanishes as investors are increasingly able to meet each other quickly, as with the case of competing marketmakers. In contrast, however, more frequent contact between investors and a monopolistic marketmaker actually widens spreads, because of the investors' poorer outside options. Specifically, an investor's threat to find a counterparty himself is less credible if the marketmaker has already executed most of the efficient trades, making it harder to find potential counterparties.

Our results regarding the role of investors' searches for each other on dealer spreads are similar in spirit to those of Gehrig (1993) and Yavaş (1996), who consider monopolistic marketmaking in one-period models.[4] We show that dynamics have an important effect on agents' bargaining positions, and thus asset prices, bid-ask spreads, and investments in marketmaking capacity. Rubinstein and Wolinsky (1987) study the complementary effects of marketmaker inventory and consignment agreements in a dynamic search model.

We consider marketmakers' choices of search intensity, and the social effi-

marketmakers may result in lower inventory costs.

[4]See also Bhattacharya and Hagerty (1987) who introduce dealers into the Diamond (1982) model, and Moresi (1991) who considers intermediation in a search model in which buyers and sellers exit the market after they trade.

ciency of these choices. A monopolistic marketmaker imposes additional "networking losses" on investors because his intermediation renders less valuable the opportunity of investors to trade directly with each other. A monopolistic marketmaker thus provides more intermediation than is socially efficient. Competitive marketmakers may provide even more intermediation, as they do not consider, in their allocation of resources to search, the effect that their intermediation has on the equilibrium allocation of assets among investors.[5]

1 Model

We fix a probability space $(\Omega, \mathcal{F}, Pr)$ and a filtration $\{\mathcal{F}_t : t \geq 0\}$ of sub-σ-algebras satisfying the usual conditions, as defined by Protter (1990). The filtration represents the resolution over time of information commonly available to agents.

Two kinds of agents, investors and marketmakers, consume a single non-storable consumption good that is used as a numeraire. All agents are risk-neutral and infinitely lived, with time preferences determined by a constant discount rate $r > 0$. Marketmakers hold no inventory and maximize profits.

Investors have access to a risk-free bank account with interest rate r and to an OTC market for a "consol," meaning an asset paying dividends at the constant rate of 1 unit of consumption per year.[6] The consol can be traded only when an investor finds another investor or a marketmaker. The associated search processes

[5]Studying endogenous search in labor markets, Mortensen (1982) and Hosios (1990) find that agents may choose inefficient search levels because they do not internalize the gains from trade realized by future trading partners. Moen (1997) shows that search markets can be efficient under certain conditions.

[6]Duffie, Gârleanu, and Pedersen (2003) consider extensions with risky securities and risk-averse investors.

are described below. The bank account can also be viewed as a liquid security that can be traded instantly. We require that the value W_t of the bank account be bounded below, ruling out Ponzi schemes.

A fraction s of investors are initially endowed with one unit of the asset. Investors can hold at most one unit of the asset and cannot shortsell. Because agents have linear utility, we can restrict attention to equilibria in which, at any given time and state of the world, an investor holds either 0 or 1 unit of the asset.

An investor is characterized by whether he owns the asset or not, and by an intrinsic type that is "high" or "low." A low-type investor, when owning the asset has a holding cost of δ per time unit. A high-type investor has no such holding cost. There are multiple interpretations of the investor types. For instance, a low-type investor may have *(i)* low liquidity (that is, a need for cash), *(ii)* high financing costs, *(iii)* hedging reasons to sell,[7] *(iv)* a relative tax disadvantage,[8] or *(v)* a lower personal use of the asset. Any investor's intrinsic type switches from low to high with intensity λ_u, and switches back with intensity λ_d. For any pair of investors, their intrinsic-type processes are assumed to be independent.

The full set of investor types is $\mathcal{T} = \{ho, hn, lo, ln\}$, with the letters "$h$" and "$l$" designating the investor's intrinsic liquidity state, as above, and with "o" or "n" indicating whether the investor owns the asset or not, respectively.

We suppose that there is a "continuum" (a non-atomic finite-measure space) of investors, and let $\mu_\sigma(t)$ denote the fraction at time t of investors of type $\sigma \in \mathcal{T}$. Because the fractions of each type of investor add to 1 at any time t,

$$\mu_{ho}(t) + \mu_{hn}(t) + \mu_{lo}(t) + \mu_{ln}(t) = 1. \tag{1}$$

[7] Duffie, Gârleanu, and Pedersen (2003) explore this interpretation in an extension with risk aversion.

[8] Dai and Rydqvist (2003) provide a tax example with potential search effects.

Because the total fraction of investors owning an asset is s,

$$\mu_{ho}(t) + \mu_{lo}(t) = s. \tag{2}$$

A pair of investors can negotiate a trade of the consol whenever they meet, for a mutually agreeable number of units of current consumption. (The determination of the terms of trade is to be addressed later.) Investors meet, however, only at random times, in a manner idealized as follows. At the successive event times of a Poisson process with some intensity parameter λ, an investor contacts another agent, chosen from the entire population "at random," meaning with a uniform distribution across the investor population.[9] Random switches in intrinsic types types are independent of the matching processes. Hence, an investor contacts an investor from a group D with intensity $\lambda\mu_D$, where μ_D is the fraction of D investors in the population. Thus, another group C of investors contact group D investors at a total rate of $\lambda\mu_C\mu_D$. Since group D investors contact C investors at the same total rate, the total meeting intensity between the two groups is $2\lambda\mu_C\mu_D$.

Marketmakers are also found through search, implying that an investor must bargain with marketmakers sequentially, as they are found. There is a unit mass of independent non-atomic marketmakers with a fixed intensity, ρ, of meeting an investor.[10] When an investor meets a marketmaker, they bargain over the terms of trade as described in the next section. Marketmakers have access to an immediately accessible inter-dealer market, on which the unload their positions, so that they have no inventory at any time.

[9]The exponential inter-contact-time distribution is natural, as it would arise from Bernoulli (independent success-failure) contact trials, with a success probability of $\lambda\Delta$ during a contact-time interval of length Δ, in the limit as Δ goes to zero. The analysis further relies on independence assumptions and an application of the law of large numbers, formalized in Duffie and Sun (2004). (See, also, Footnote 13).

[10]It would be equivalent to have a mass k of dealers with contact intensity ρ/k, for any $k > 0$.

Certain over-the-counter markets do not have marketmakers. Such markets are, of course, described by the special case of our model with $\rho = 0$. Hence, our results to follow characterize allocations and inter-investor prices both with and without marketmakers.

2 Dynamic Search Equilibrium with Competing Marketmakers

In this section, we explicitly compute the allocations and prices forming a dynamic search-and-bargaining equilibrium. In particular, we compute the price negotiated directly between investors, marketmaker's bid and ask prices, and the inter-dealer price.

In equilibrium, low-type asset owners want to sell and high-type non-owners want to buy. When two such agents meet, they bargain over the price. Similarly, when investors meet a marketmaker, they bargain over the price. An investor's bargaining position depends on his outside option, which in turn depends on the availability of other counterparties, both now and in the future, and a marketmaker's bargaining position depends on the inter-dealer price. In deriving the equilibrium, we rely on the insight from bargaining theory that trade happens instantly.[11] This allows us to derive a dynamic equilibrium in two steps. First, we derive the equilibrium masses of the different investor types. Second, we compute agents' value functions and transaction prices (taking as given the masses).

Assuming that the law of law of large numbers applies (see Duffie and Sun

[11] In general, bargaining leads to instant trade when agents do not have asymmetric information. Otherwise there can be strategic delay. In our model, it does not matter whether agents have private information about their own type for it is common knowledge that a gain from trade arises only between between agents of types lo and hn.

(2004)), the rate of change of the mass $\mu_{lo}(t)$ of low-type owners is almost surely

$$\dot{\mu}_{lo}(t) \;=\; -\left(2\lambda\mu_{hn}(t)\mu_{lo}(t) + \rho\mu_m(t)\right) - \lambda_u\mu_{lo}(t) + \lambda_d\mu_{ho}(t), \qquad (3)$$

where $\mu_m(t) = \min\{\mu_{lo}(t), \mu_{hn}(t)\}$. The first term in (3) reflects the fact that agents of type hn contact those of type lo at a total rate of $\lambda\mu_{hn}(t)\mu_{lo}(t)$, while agents of type lo contact those of type hn at the same total rate $\lambda\mu_{hn}(t)\mu_{lo}(t)$. At both of these types of encounters, the agent of type lo becomes one of type ln. This implies a total rate of reduction of mass due to these encounters of $2\lambda\mu_{hn}(t)\mu_{lo}(t)$. Similarly, investors of type lo meet marketmakers with a total contact intensity of $\rho\mu_{lo}(t)$. If $\mu_{lo}(t) \leq \mu_{hn}(t)$ then all these meetings lead to trade, and the lo agent becomes a ln agent, resulting in a reduction of μ_{lo} of $\rho\mu_{lo}(t)$. If $\mu_{lo}(t) > \mu_{hn}(t)$, then not all these meetings result in trade. This is because marketmakers buy from lo investors and sell to hn investors, and, in equilibrium, the total intensity of selling must equal the intensity of buying. Marketmakers meet lo-investors with total intensity $\rho\mu_{lo}$ and hn-investors with total intensity $\rho\mu_{hn}$, and, therefore, investors on the "long side" of the market are rationed. In particular, if $\mu_{lo}(t) > \mu_{hn}(t)$ then lo agents trade with marketmakers only at the intensity $\rho\mu_{hn}$. In equilibrium this rationing can be the outcome of bargaining because the marketmaker's reservation value (that is, the inter-dealer price) is equal to the reservation value of the lo-investor.

Finally, the term $\lambda_u\mu_{lo}(t)$ reflects the migration of owners from low to high intrinsic types, and the last term $\lambda_d\mu_{ho}(t)$ reflects owners' change from high to low intrinsic types.

The rate of change of the other investor-type masses are,

$$\dot{\mu}_{hn}(t) = -\left(2\lambda\mu_{hn}(t)\mu_{lo}(t) + \rho\mu_m(t)\right) + \lambda_u\mu_{ln}(t) - \lambda_d\mu_{hn}(t) \quad (4)$$

$$\dot{\mu}_{ho}(t) = \left(2\lambda\mu_{hn}(t)\mu_{lo}(t) + \rho\mu_m(t)\right) + \lambda_u\mu_{lo}(t) - \lambda_d\mu_{ho}(t) \quad (5)$$

$$\dot{\mu}_{ln}(t) = \left(2\lambda\mu_{hn}(t)\mu_{lo}(t) + \rho\mu_m(t)\right) - \lambda_u\mu_{ln}(t) + \lambda_d\mu_{hn}(t). \quad (6)$$

As in (3), the first terms reflect the result of trade, and the last two terms are the result of intrinsic-type changes.

In most of the paper we focus on stationary equilibria, that is, equilibria in which the masses are constant. In our welfare analysis, however, it is more natural to take the initial masses as given, and, therefore, we develop some results with any initial mass distribution. The following proposition asserts the existence, uniqueness, and stability of the steady state.

Proposition 1 *There exists a unique constant steady-state solution to (1)–(6). From any initial condition $\mu(0) \in [0,1]^4$ satisfying (1) and (2), the unique solution $\mu(t)$ to (3)–(6) converges to the steady state as $t \to \infty$.*

A particular agent's type process $\{\sigma(t) : -\infty < t < +\infty\}$ is, in steady-state, a 4-state Markov chain with state space $\mathcal{T}$, and with constant switching intensities determined in the obvious way[12] by the steady-state population masses μ and the intensities λ, λ_u, and λ_d. The unique stationary distribution of any agent's type process coincides with the cross-sectional distribution μ of types characterized[13] in Proposition 1.

[12]For example, the transition intensity from state lo to state ho is λ_u, the transition intensity from state lo to state ln is $2\lambda\mu_{hn}$, and so on, for the 4×3 switching intensities.

[13]This is a result of the law of large numbers, in the form of Theorem C of Sun (2000), which provides the construction of our probability space $(\Omega, \mathcal{F}, Pr)$ and agent space $[0,1]$, with an appropriate σ-algebra making $\Omega \times [0,1]$ into what Sun calls a "rich space," with the properties that: *(i)* for each individual agent in $[0,1]$, the agent's type process is indeed a Markov chain in $\mathcal{T}$ with the specified generator, *(ii)* the unconditional probability distribution of the agents' type is always

Turning to the determination of an equilibrium transaction prices, we first conjecture, and verify shortly, a natural steady-state equilibrium utility for remaining lifetime consumption.

With these equilibrium masses, we will determine the price P negotiated directly between lo and hn investors, the "bid" price B at which investors sell to marketmakers, the "ask" price A at which investors buy from marketmakers, and the inter-dealer price. For this, we use dynamic programming, by first computing an investor's utility at time t for remaining lifetime consumption. For a particular agent this "value function" depends, naturally, only on the agent's current type $\sigma(t) \in \mathcal{T}$, the current wealth W_t in his bank account, and time. Because of risk neutrality, the value function has the form $W_t + V_{\sigma(t)}(t)$. Because any budget-feasible consumption withdrawals from liquid wealth is optimal, we simply assume that agents adjust their consumption so that $W_t = 0$ for all t. As shown in the appendix, the value functions satisfy:

$$\begin{aligned}
\dot{V}_{lo} &= rV_{lo} - \lambda_u(V_{ho} - V_{lo}) - 2\lambda\mu_{hn}(P + V_{ln} - V_{lo}) - \rho(B + V_{ln} - V_{lo}) - (1 - \delta) \\
\dot{V}_{ln} &= rV_{ln} - \lambda_u(V_{hn} - V_{ln}) \\
\dot{V}_{ho} &= rV_{ho} - \lambda_d(V_{lo} - V_{ho}) - 1 \\
\dot{V}_{hn} &= rV_{hn} - \lambda_d(V_{ln} - V_{hn}) - 2\lambda\mu_{ho}(V_{ho} - V_{hn} - P) - \rho(V_{ho} - V_{hn} - A),
\end{aligned} \tag{7}$$

where the value functions (V_σ), prices (P, A, B), and masses (μ_σ), depend on time unless the initial masses are the steady-state ones.

These value functions imply that an lo-investor benefits from a sale at any price greater than $V_{lo} - V_{ln}$, and an hn-investor will benefit from a purchase at

the steady-state distribution μ on $\mathcal{T}$ given by Proposition 1, *(iii)* agents' type transitions are almost everywhere pair-wise independent, and *(iv)* the cross-sectional distribution of types is also given by μ, almost surely, at each time t.

any price smaller than $V_{ho} - V_{hn}$. Bargaining between the investors leads to a price between these two values. Specifically, Nash (1950) bargaining with seller bargaining power of $q \in [0, 1]$ yields

$$P = (V_{lo} - V_{ln})(1 - q) + (V_{ho} - V_{hn})q. \tag{8}$$

This is also the outcome of the simultaneous-offer bargaining game described in Kreps (1990), and of the alternating-offers bargaining game in Duffie, Gârleanu, and Pedersen (2003).

Similarly, the bid and ask prices are determined through a bargaining encounter between investors and marketmakers in which a marketmaker's outside option is to trade in the interdealer market at a price of M. Marketmakers have a fraction, $z \in [0, 1]$, of the bargaining power when facing an investor. Hence, a marketmaker buys from an investor at the bid price B, and sells at the ask price A, determined by

$$A = (V_{ho} - V_{hn})z + M(1 - z) \tag{9}$$

$$B = (V_{lo} - V_{ln})z + M(1 - z). \tag{10}$$

As discussed above, in equilibrium, the marketmakers and the investors on the long side of the market must be indifferent to trading. Hence, if $\mu_{lo} \leq \mu_{hn}$, marketmakers meet more potential buyers than sellers. The inter-dealer price, M, is therefore equal to the ask price, A, and to any buyer's reservation value, $V_{ho} - V_{hn}$. Similarly, if $\mu_{lo} > \mu_{hn}$, then $M = B = V_{lo} - V_{ln}$.

In steady state, it is easy to see which side of the market is rationed because

the steady-state fraction of high-type agents is $\lambda_u(\lambda_d+\lambda_u)^{-1}$, so we have

$$\mu_{hn} + (s - \mu_{lo}) = \frac{\lambda_u}{\lambda_d + \lambda_u}.$$

Hence, $\mu_{lo} < \mu_{hn}$ in steady state if and only if the following condition is satisfied.

Condition 1 $s < \lambda_u/(\lambda_u + \lambda_d)$.

The equations for prices and value functions can be solved explicitly. Condition 1 seems the natural case, and the solution in that case is given by the following theorem; the complementary case is treated in the appendix.

Theorem 2 *For any given initial mass distribution $\mu(0)$, there exists a subgame-perfect Nash equilibrium. There is a unique steady-state equilibrium. Under Condition 1, the ask, bid, and inter-investor prices are*

$$A = \frac{1}{r} - \frac{\delta}{r}\frac{\lambda_d + 2\lambda\mu_{lo}(1-q)}{r + \lambda_d + 2\lambda\mu_{lo}(1-q) + \lambda_u + 2\lambda\mu_{hn}q + \rho(1-z)} \tag{11}$$

$$B = \frac{1}{r} - \frac{\delta}{r}\frac{zr + \lambda_d + 2\lambda\mu_{lo}(1-q)}{r + \lambda_d + 2\lambda\mu_{lo}(1-q) + \lambda_u + 2\lambda\mu_{hn}q + \rho(1-z)} \tag{12}$$

$$P = \frac{1}{r} - \frac{\delta}{r}\frac{(1-q)r + \lambda_d + 2\lambda\mu_{lo}(1-q)}{r + \lambda_d + 2\lambda\mu_{lo}(1-q) + \lambda_u + 2\lambda\mu_{hn}q + \rho(1-z)}. \tag{13}$$

These explicit prices are intuitive. Each price is the present value, $1/r$, of dividends, reduced by an illiquidity discount. All of these prices decrease in the bargaining power, z, of the marketmaker, because a higher z makes trading more costly for investors. The prices increase, however, in the ease of meeting a marketmaker (ρ) and in the ease of finding another investor (λ), provided ρ and λ are large enough. The interesting effects of high search intensities are discussed in detail in Section 4.

From Theorem 2, the bid-ask spread $(A - B)$ is increasing in the marketmaker's bargaining power z. The bid-ask spread is decreasing in λ, since a high λ means that an investor can easily find a counterparty himself, which improves his bargaining position. The bid-ask spread is also decreasing in ρ, provided $z < 1$ and ρ is sufficiently large. A higher ρ implies that an investor can quickly find another marketmaker, and this "sequential competition" improves his bargaining position. If $z = 1$, however, then the bid-ask spread is increasing in ρ. The case of $z = 1$ is best interpreted as a monopolistic marketmaker as we show in the next section. These comparative-statics results can be derived from the price equations (11)–(13) and from Equation (A.2), which characterizes the steady-state investor masses.

3 Monopolistic Marketmaking

We assume here that investors can trade with the monopolistic marketmaker only when they meet one of the marketmaker's non-atomic "dealers." There is a unit mass of such dealers who contact potential investors randomly and pair-wise independently, letting ρ be the intensity with which a dealer contacts a given agent.

Dealers instantly balance their positions with their marketmaking firm, which, on the whole, holds no inventory. When an investor meets a dealer, the dealer is assumed to have all of the bargaining power since the marketmaker's profit is not affected by any one infinitesimal trade. Hence, the dealer quotes an ask price, A, and a bid price, B, that are, respectively, a buyer's and a seller's reservation value.

With these assumptions, the equilibrium is computed as in Section 2. The masses are determined by (3)–(6) and the prices are given by Theorem 2 with $z = 1$. In equilibrium, $B \leq P \leq A$.

It might seem surprising that a single monopolistic marketmaker is equivalent for pricing purposes to many "competing" non-atomic marketmakers with full bargaining power ($z = 1$). The result follows from the fact that a search economy is inherently un-competitive, in that each time agents meet, a bilateral bargaining relationship obtains. With many non-atomic marketmakers it is, however, more natural to assume that $z < 1$, and, hence, this difference in marketmaker bargaining power distinguishes the two kinds of intermediation. The distinction between monopolistic and competitive marketmakers is clearer when search intensities are endogenized in Section 7.

4 Fast Search Leads to Competitive Prices?

A competitive Walrasian equilibrium is characterized by a single price process at which agents may buy and sell *instantly*, such that supply equals demand in each state and at every point in time. A Walrasian allocation is efficient and all assets are held by agents of high type, if there are enough such agents,[14] which is the case in steady state if $s < \lambda_u/(\lambda_u + \lambda_d)$. If $s > \lambda_u/(\lambda_u + \lambda_d)$, all high-type agents own assets, and the rest of the assets are held by low-type investors. Finally, if $s = \lambda_u/(\lambda_u + \lambda_d)$, the number of sellers is equal to the number of buyers.

[14] The quantity of such agents can be thought, for instance, as the capacity for taking a certain kind of risk.

In the former case, the unique Walras equilibrium has agent masses

$$\begin{aligned}
\mu^*_{ho} &= s \\
\mu^*_{hn} &= \frac{\lambda_u}{\lambda_u+\lambda_d} - s \\
\mu^*_{lo} &= 0 \\
\mu^*_{ln} &= \frac{\lambda_d}{\lambda_u+\lambda_d},
\end{aligned} \tag{14}$$

and the Walrasian price is

$$P^* = E_t\left[\int_0^\infty e^{-rs}\,ds\right] = \frac{1}{r}.$$

The Walras equilibrium price, a version of what is sometimes called the "Gordon dividend growth model" of valuation, is the reservation value of holding the asset forever for a hypothetical investor who is always of high type.

In case of $s > \lambda_u/(\lambda_u+\lambda_d)$, the masses are determined similarly and since the marginal investor has low liquidity, the Walrasian price is the reservation value of holding the asset indefinitely for a hypothetical agent who is permamently of low type (that is, $P^* = (1-\delta)/r$). If $s = \lambda_u/(\lambda_u+\lambda_d)$, then any price P^* between $1/r$ and $(1-\delta)/r$ is a Walrasian equilibrium.

Clearly, fast search by either investors or marketmakers implies that allocations approach the efficient allocations, μ^*, prevailing in a Walrasian market. The following theorem further determines the circumstances under which prices approach the competitive Walrasian prices, P^*.

Theorem 3 *Let* $(\lambda^k, \rho^k, \mu^k, B^k, A^k, P^k)$ *be a sequence of stationary search equilibria.*

1. [Fast investors.] *If* $\lambda^k \to \infty$, (ρ^k) *is any sequence, and* $0 < q < 1$ *then*

$\mu^k \to \mu^*$, *and* B^k, A^k, *and* P^k *converge to the same price, which is Walrasian.*

2. [Fast competing marketmakers.] *If* $\rho^k \to \infty$, (λ^k) *is any sequence, and* $z < 1$ *then* $\mu^k \to \mu^*$, *and* B^k, A^k, *and* P^k *converge to the same price, which is Walrasian.*

3. [Fast monopolistic marketmaker.] *If* $\lambda^k = \lambda$ *is constant,* $\rho^k \to \infty$ *is an increasing sequence, and* $z = 1$, *then* $\mu^k \to \mu^*$ *and the bid-ask spread,* $A^k - B^k$, *is increasing.*

Part one shows that prices become competitive and that the bid-ask spread approaches zero as investors find *each other* more quickly, regardless of the nature of intermediation. In other words, the investors' search alternative forces the marketmakers to offer relatively competitive prices, consistent with the evidence of Lamoureux and Schnitzlein (1997).[15]

Part two shows that fast intermediation by competing marketmakers also leads to competitive prices and vanishing bid-ask spreads. This may seem surprising, given that an investor trades with the first encountered marketmaker, and this marketmaker could have almost all bargaining power (z close to 1). As ρ increases, however, the investor's outside option when bargaining with a marketmaker improves, because he can more easily meet another marketmaker, and this sequential competition ultimately results in competitive prices.

Part three shows that fast intermediation by a monopolistic marketmaker does not lead to competitive prices. In fact, the bid-ask spread *widens* as intermedia-

[15]This result holds, under certain conditions, even if the monopolistic marketmaker can be approached instantly ("$\rho = +\infty$"). In this case, for any finite λ, *all* trades are done using the marketmaker, but as the investors' outside options improve, even a monopolistic marketmaker needs to quote competitive prices.

tion by marketmakers increases. This is because an investor's potential "threat" to search for a direct trade with another investor becomes increasingly less persuasive, since the mass of investors with whom there are gains from trade shrinks.

Contrary to our result, Rubinstein and Wolinsky (1985) find that their bargaining equilibrium (without intermediaries) does *not* converge to the competitive equilibrium as trading frictions approach zero. Gale (1987) argues that this failure is due to the fact that the total mass of agents entering their economy is infinite, which makes the competitive equilibrium of the total economy undefined. Gale (1987) shows that if the total mass of agents is finite, then the economy (which is not stationary) is Walrasian in the limit. He suggests that, when considering stationary economies, one should compare the bargaining prices to those of a "flow equilibrium" rather than a "stock equilibrium." Our model has a natural determination of steady-state masses, even though no agent enters the economy. This is accomplished by considering agents whose types change over time.[16] We are able to reconcile a steady-state economy with convergence to Walrasian outcomes in both a flow and stock sense, both for allocations and for prices, and by increasing both investor search and marketmaker search.[17]

5 Numerical Example

We illustrate some of the search effects on asset pricing and marketmaking with a numerical example. Figure 1 shows the marketmakers' bid (B), and ask (A)

[16]Gale (1986a), Gale (1986b), and McLennan and Sonnenschein (1991) show that a bargaining game implements Walrasian outcomes in the limiting case with no frictions (that is, no discounting) in much richer settings for preferences and goods. See also Binmore and Herrero (1988).

[17]Other important differences between our framework and that of Rubinstein and Wolinsky (1985) are that we accommodate repeated trade, and that we diminish search frictions explicitly through λ rather than implicitly through the discount rate. See Bester (1988, 1989) for the importance of diminishing search frictions directly.

prices, as well as the inter-investor price (P). These prices are plotted as functions of the intensity, ρ, of meeting dealers. The top panel deals with the case of competing marketmakers with bargaining power $z = 0.8$, whereas the bottom panel treats a monopolistic marketmaker ($z = 1$). The parameters underlying these graphs are as follows. First, $\lambda_d = 0.1$ and $\lambda_u = 1$, which implies that an agent is of high liquidity type 91% of the time. An investor finds other investors on average every two weeks, that is, $\lambda = 26$, and selling investors have bargaining power $q = 0.5$. The supply is $s = 0.8$, and the interest rate is $r = 0.05$

Since allocations become more efficient as ρ increases, in both cases, all prices increase with ρ. Interestingly, in the case of competing marketmakers ($z = 0.8$), the price increases to the Walrasian price $1/r = 20$ and the bid-ask spread decreases to zero. In the case of a monopolist marketmaker ($z = 1$), on the other hand, the prices are bounded away from $1/r = 20$, and the bid-ask spread is increasing in ρ.

The intuition for this difference is as follows. When the dealers' contact intensities increase, they execute more trades. Investors then find it more difficult to contact other investors with whom to trade. If dealers have all of the bargaining power, this leads to wider spreads. If dealers don't have all of the bargaining power, however, then higher marketmaker intensity leads to a narrowing of the spread, because an investor has an improved threat of waiting to trade with the next encountered marketmaker.

6 Heterogeneous Investors

So far, we have assumed that investors are homogeneous with respect to the speed with which they find counterparties. In certain OTC markets, however, some in-

49.2

49.4

PSfrag replacements

49.8

500

50

1000

48.65

48.7

48.75

Monopolistic Dealer, $z = 1$

Competitive Dealers, $z = 0.8$

ρ

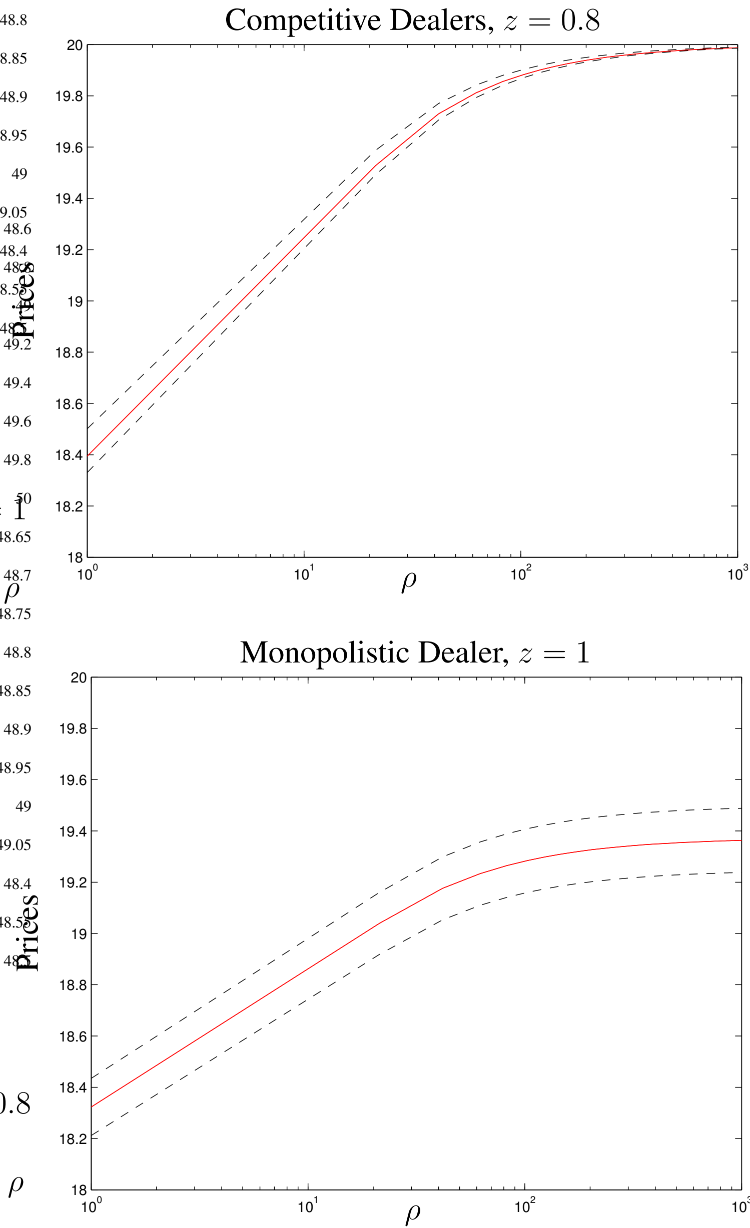

Figure 1: The solid line shows the price P used when investors trade with each other; the dashed lines show the bid (B) and ask (A) prices used when investors trade with a marketmaker. The prices are functions of the intensity (ρ) with which an investor meets a dealer, which is plotted on a logarithmic scale. The bargaining power of the marketmaker is $z = 0.8$ in the left panel, and $z = 1$ in the right panel.

vestors are more sophisticated than others, in the sense that they have faster and easier access to counterparties. To capture this effect, we assume that there are two different investor classes, ”sophisticated,” of total mass μ^s, and “unsophisticated,” of mass $1-\mu^s$. We assume that sophisticated investors meet marketmakers with an intensity ρ^s, while unsophisticated investors meet marketmakers at intensity ρ^u, where $\rho^u < \rho^s$. We assume here that investors cannot trade directly with each other, that is, $\lambda = 0$. If this assumption is relaxed, and investors are able to find each other (possibly with type-dependent speeds), then the nature of the equilibrium that we will describe would change for certain parameters. In particular, sophisticated investors would, under certain conditions, profit from executing as many trades as possible, and would start acting like marketmakers. This interesting effect is beyond the scope if this paper; we focus on how marketmakers react to differences in investor sophistication.

An investor’s type is observable to the marketmakers, who have bargaining power $z < 1$. When a sophisticated investor meets a marketmaker, the outcome of their bargaining is a bid price of B^s or an ask price of A^s.

When an unsophisticated investor needs to buy or sell, locating a marketmaker takes more time. This results in higher expected holding costs associated with illiquidity and, importantly, implies a poor bargaining position. Hence, unsophisticated investors receive different bid and ask prices, which we denote by B^u and A^u, respectively.

When the supply of shares is so low that the sophisticated investors are “marginal” buyers, then all unsophisticated investors optimally stay out of the market, that is, they never buy any shares. Similarly, when the supply of shares is large, sophisticated investors are marginal sellers, and all unsophisticated investors hold a share each, never selling. With an intermediate supply, all investors trade, but the unso-

phisticated investors trade at a larger spread.

The following theorem characterizes the most important properties of the equilibrium with heterogeneous investors; a full characterization is in the appendix.

Theorem 4 *If $s < \mu^s \frac{\lambda_u}{\lambda_u+\lambda_d}$ or $s > 1 - \mu^s \frac{\lambda_d}{\lambda_u+\lambda_d}$ then unsophisticated investors do not trade. Otherwise, all investors trade, and marketmakers quote a larger bid-ask spread to unsophisticated investors than to sophisticated investors. That is, $A^u - B^u > A^s - B^s$. In particular, an agent who meets a marketmaker with intensity ρ faces a bid-ask of*

$$A - B = \frac{z\delta}{r + \lambda_u + \lambda_d + \rho(1-z)}\,. \tag{15}$$

7 Endogenous Search and Welfare

Here, we investigate the search intensities that marketmakers would optimally choose in the two cases considered above: a single monopolistic marketmaker and non-atomic competing marketmakers. We illustrate how marketmakers' choices of search intensities depend on: *(i)* a marketmaker's personal influence on the equilibrium allocations of assets, and *(ii)* a marketmaker's bargaining power. We take investors' search intensities as given. Considering the interactions arising if both investors and intermediaries choose search intensities would be an interesting issue for future research.[18]

Because the marketmakers' search intensities, collectively, affect the masses μ of investor types, it is natural to take as given the initial masses, $\mu(0)$, of investors, rather than to compare based on the different steady-state masses corresponding

[18]Related to this, Pagano (1989) considers a one-period model in which investors choose between searching for a counterparty and trading on a centralized market.

to different choices of search intensities. Hence, in this section, we are not relying on a steady-state analysis.

We assume that a marketmaker chooses one search intensity and abides by it. This assumption is convenient, and can be motivated by interpreting the search intensity as based on a technology that is difficult to change. A full dynamic analysis of the optimal control of marketmaking intensities with small switching costs would be interesting, but seems difficult. We merely assume that marketmakers choose ρ so as to maximize the present value, using their discount rate r, of future marketmaking spreads, net of the rate $\Gamma(\rho)$ of technology costs, where $\Gamma : [0,\infty) \to [0,\infty)$ is assumed for technical convenience to be continuously differentiable, strictly convex, with $\Gamma(0) = 0$, $\Gamma'(0) = 0$, and $\lim_{\rho\to\infty} \Gamma'(\rho) = \infty$.

The marketmaker's trading profit, per unit of time, is the product of the volume of trade, $\rho\mu_m$, and the bid-ask spread, $A-B$. Hence, a monopolistic marketmaker who searches with an intensity of ρ has an initial valuation of

$$\pi^M(\rho) = E\left[\int_0^\infty \rho\mu_m(t,\rho)\left(A(t,\rho) - B(t,\rho)\right)e^{-rt}\,dt\right] - \frac{\Gamma(\rho)}{r}, \tag{16}$$

where $\mu_m = \min\{\mu_{lo}, \mu_{hn}\}$, and where we are using the obvious notation to indicate dependence of the solution on ρ and t.

Any one non-atomic marketmaker does not influence the equilibrium masses of investors, and therefore values his profit at

$$\pi^C(\rho) = \rho\, E\left[\int_0^\infty \mu_m(t)\left(A(t) - B(t)\right)e^{-rt}\,dt\right] - \frac{\Gamma(\rho)}{r}.$$

An equilibrium intensity, ρ^C, for non-atomic marketmakers is a solution to the

first-order condition

$$\Gamma'(\rho^C) = rE\left[\int_0^\infty \mu_m(t,\rho^C)\left(A(t,\rho^C) - B(t,\rho^C)\right) e^{-rt}\, dt\right]. \tag{17}$$

The following theorem characterizes equilibrium search intensities in the case of "patient" marketmakers.

Theorem 5 *There exists a marketmaking intensity ρ^M that maximizes $\pi^M(\rho)$. There exists $\bar{r} > 0$ such that, for all $r < \bar{r}$ and for each $z \in [0,1]$, unique number $\rho^C(z)$ solves (17). Moreover, $\rho^C(0) = 0$, $\rho^C(z)$ is increasing in z, and $\rho^C(1)$ is larger than any solution, ρ^M, to the monopolist's problem.*

In addition to providing the existence of equilibrium search intensities, this result establishes that: *(i)* competing marketmakers provide more marketmaking services if they can capture a higher proportion of the gains from trade, and *(ii)* competing marketmakers with full bargaining power provide more marketmaking services than a monopolistic marketmaker, since they do not internalize the consequences of their search on the masses of investor types.

To consider the welfare implications of marketmaking in our search economy, we adopt a notion of "social welfare," the sum of the utilities of investors and marketmakers. This can be interpreted as the total investor utility in the case in which the marketmaker profits are redistributed to investors, for instance through share holdings. With our form of linear preferences, maximizing social welfare is a meaningful concept in that it is equivalent to requiring that utilities cannot be Pareto improved by changing allocations and by making initial consumption transfers.[19] By "investor welfare," we mean the total of investors' utilities, as-

[19]Also, this "utilitarian" social welfare function can be justified by considering the utility of an agent "behind the veil of ignorance," not knowing what type of agent he will become.

suming that the marketmaker profits are not redistributed to investors. We take "marketmaker welfare" to be the total valuation of marketmaking profits, net of the cost of intermediation.

In our risk-neutral framework, welfare losses are easily quantified. The total "social loss" is the cost $\Gamma(\rho)$ of intermediation plus the present value of the stream $\delta\mu_{lo}(t), t \geq 0$, of dividends wasted through mis-allocation. At a given marketmaking intensity ρ, this leaves the social welfare

$$w^S(\rho) = E\left[\int_0^\infty (s - \delta\mu_{lo}(t))\, e^{-rt}\, dt\right] - \frac{\Gamma(\rho)}{r}.$$

Investor welfare is, similarly,

$$w^I(\rho) = E\left[\int_0^\infty (s - \delta\mu_{lo}(t,\rho) - \rho\mu_m(t,\rho)(A(t,\rho) - B(t,\rho)))\, e^{-rt}\, dt\right],$$

and the marketmakers' welfare is

$$w^M(\rho) = E\left[\int_0^\infty \rho\mu_m(t,\rho)(A(t,\rho) - B(t,\rho))e^{-rt}\, dt\right] - \frac{\Gamma(\rho)}{r}.$$

We consider first the case of monopolistic marketmaking. We let ρ^M be the level of intermediation optimally chosen by the marketmaker, and ρ^S be the socially optimal level of intermediation. The relation between the monopolistic marketmaker's chosen level ρ^M of intensity and the socially optimal intensity ρ^S is characterized in the following theorem.

Theorem 6 *Let $z = 1$. (i) If investors cannot meet directly, that is, $\lambda = 0$, then the investor welfare $w^I(\rho)$ is independent of ρ, and a monopolistic marketmaker provides the socially optimal level ρ^S of intermediation, that is, $\rho^M = \rho^S$.*
(ii) If $\lambda > 0$, then $w^I(\rho)$ decreases in ρ, and the monopolistic marketmaker over-

invests in intermediation, that is, $\rho^M > \rho^S$, provided q is 0 or 1.

The point of this result is that if investors cannot search, then their utilities do not depend on the level of intermediation because the monopolist extracts all gains from trade. In this case, because the monopolist gets all social benefits from providing intermediation and bears all the costs, he chooses the socially optimal level.

If, on the other hand, investors can trade directly with each other, then the marketmaker may exploit the opportunity to invest in additional search for trades in order to reduce the opportunities of investors to trade directly with each other. Therefore, investor welfare decreases with ρ. Consequently, the marketmaker's marginal benefit from intermediation is larger than the social benefit, so there is too much intermediation.[20]

We now turn to the case of non-atomic (competing) marketmakers. We saw above that the equilibrium level of intermediation of a non-atomic marketmaker depends critically on its bargaining power. With no bargaining power, such a marketmaker provides no intermediation. With complete bargaining power, they search more than a monopolistic marketmaker would.

A government may sometimes be able to affect intermediaries' market power, for instance through the enforcement of regulation (DeMarzo, Fishman, and Hagerty (2000)). Hence, we consider the following questions: How much marketmaker market power is socially optimal? How much market power would the intermediaries like to have? Would investors want that marketmakers to have some market power? These questions are answered in the following theorem, in which we let z^I, z^S, and z^M denote the marketmaker bargaining power that would be chosen

[20] If $0 < q < 1$, then increasing ρ has the additional effect of changing the relative strength of investors' bargaining positions with the marketmaker, because it changes their outside options, which complicates the calculations.

by, respectively, the investors, a social-welfare maximizing planner, and marketmakers.

Theorem 7 *It holds that* $z^I > 0$. *There is some* $\bar{r} > 0$ *such that, provided* $r < \bar{r}$, *we have* $z^I < z^S \leq z^M = 1$.

Investors in our model would prefer to enter a market in which non-atomic marketmakers have some market power, because this gives marketmakers an incentive to provide intermediation. The efficient level of intermediation is achieved with a higher market power to marketmakers. Marketmakers themselves prefer to have full bargaining power.

8 Empirical Implications

This paper lays out a theory of asset pricing and marketmaking based on search and bargaining. We show how search-based inefficiencies affect prices through equilibrium allocations and through the effect of search on agents' bargaining position, that is, their outside options based on their ability to trade with other investors or marketmakers.

Consider, for example, the OTC market for interest-rate swaps, which, according to the British Bankers Association has open positions totally roughly $100 trillion dollars. Customers rarely have material private information about the current level of the underlying interest rates, so standard information-based explanations of bid-ask spreads are not compelling in this market. Instead, a "sales trader" sets spreads based on a customer's (perceived) outside option, and would rarely fear that the customer has superior information about the underlying interest rates. The customer's outside option depends on how easily he can find a counterparty

himself (proxied by λ in our model), and how easily he can access other banks (proxied by ρ in our model). To trade OTC derivatives with a bank one needs, among other things, an account and a credit clearing. Small investors often only have an account with one or few banks, implying that such investors have lower search options. Hence, a testable implication of our search framework is that (small) investors with lower search options receive less competitive prices. We note that these investors are less likely to be informed, so traditional information-based models of spreads (for example, Glosten and Milgrom (1985)), applied to this market, would have the opposite prediction.

The model that we present here can also be viewed as one of imperfect competition, for example in a specialist-based equity markets. In particular, the model shows that even a monopolistic marketmaker may have a tight bid-ask spread if investors can easily trade directly with each other (that is, have a high λ). This resembles situations at the New York Stock Exchange (NYSE) in which there is a single specialist for each stock, but with floor brokers who can find each other relatively easily and trade directly, and with outside brokers who can "find each other" and trade around the specialist by submitting limit orders. On Nasdaq, however, a "phone market" with several dealers for each stock, it can be difficult for investors to find each other directly. Before the reforms in 1994, 1995, and 1997, it was difficult for investors to compete with the Nasdaq marketmakers through limit orders.[21] This may help explain why spreads were higher on Nasdaq than on NYSE (Huang and Stoll (1996)). Consistent with this view, Barclay, Christie, Harris, Kandel, and Schultz (1999) find that the "Securities and Exchange Commission began implementing reforms that would permit the public to compete directly with Nasdaq dealers by submitting binding limit orders ... Our

[21] See Barclay, Christie, Harris, Kandel, and Schultz (1999) and references therein.

results indicate that quoted and effective spreads fell dramatically."

The competition faced by marketmakers from direct trade between investors can perhaps be gauged by the *participation rate* of marketmakers, that is, the fraction of trades that are intermediated by a marketmaker. Our model suggest that, with equal marketmaker availability and stock characteristics, stocks with higher participation rates are characterized by lower search intensity (λ) and, hence, higher bid-ask spreads. On Nasdaq, the participation rate was once large relative to the NYSE, which had a participation rate of between 18.8% and 24.2% in the 1990s (New York Stock Exchange (2001)). At that time, the NYSE may well have covered stocks whose investors had higher direct contact rates (λ) than those covered, on average, by Nasdaq.

A Appendix: Proofs

Proof of Proposition 1: Start by letting

$$y = \frac{\lambda_u}{\lambda_u + \lambda_d}$$

and assume that $y > s$. The case $y \leq s$ can be treated analogously. Setting the right-hand side of Equation (3) to zero and substituting all components of μ other than μ_{lo} in terms of μ_{lo} from Equations (1) and (2) and from $\mu_{lo} + \mu_{ln} = \lambda_d(\lambda_d + \lambda_u)^{-1} = 1 - y$, we obtain the quadratic equation

$$Q(\mu_{lo}) = 0, \tag{A.1}$$

where

$$Q(x) = 2\lambda x^2 + (2\lambda(y-s) + \rho + \lambda_u + \lambda_d)x - \lambda_d s. \tag{A.2}$$

It is immediate that Q has a negative root (since $Q(0) < 0$) and has a root in the interval $(0, 1)$ (since $Q(1) > 0$).

Since μ_{lo} is the largest and positive root of a quadratic with positive leading coefficient and with a negative root, in order to show that $\mu_{lo} < \eta$ for some $\eta > 0$ it suffices to show that $Q(\eta) > 0$. Thus, in order that $\mu_{ho} > 0$ (for, clearly, $\mu_{ho} < 1$), it is sufficient that $Q(s) > 0$, which is true, since

$$Q(s) = 2\lambda s^2 + (\lambda_u + 2\lambda(y-s) + \rho)s.$$

Similarly, $\mu_{ln} > 0$ if $Q(1-y) > 0$, which holds because

$$Q(1-y) = 2\lambda(1-y)^2 + (2\lambda(y-s) + \rho)(1-y) + \lambda_d(1-s).$$

Finally, since $\mu_{hn} = y - s + \mu_{lo}$, it is immediate that $\mu_{hn} > 0$.

We present a sketch of a proof of the claim that, from any admissible initial condition $\mu(0)$ the system converges to the steady-state μ.

Because of the two restrictions (1) and (2), the system is reduced to two equations, which can be thought of as equations in the unknowns $\mu_{lo}(t)$ and $\mu_l(t)$, where $\mu_l(t) = \mu_{lo}(t) + \mu_{ln}(t)$. The equation for $\mu_l(t)$ does not depend on $\mu_{lo}(t)$, and admits the simple solution:

$$\mu_l(t) = \mu_l(0)e^{-(\lambda_d+\lambda_u)t} + \frac{\lambda_d}{(\lambda_d + \lambda_u)}(1 - e^{-(\lambda_d+\lambda_u)t}).$$

Define the function

$$G(w,x) = -2\lambda x^2 - (\lambda_u + \lambda_d + 2\lambda(1-s-w)+\rho)x + \rho\max\{0, s+w-1\} + \lambda_d s$$

and note that μ_{lo} satisfies

$$\dot{\mu}_{lo}(t) = G(\mu_l(t), \mu_{lo}(t)).$$

The claim is proved by the steps:

1. Choose t_1 high enough that $s + \mu_l(t) - 1$ does not change sign for $t > t_1$.

2. Show that $\mu_{lo}(t)$ stays in $(0,1)$ for all t, by verifying that $G(w,0) > 0$ and $G(w,1) < 0$.

3. Choose t_2 ($\geq t_1$) high enough that $\mu_l(t)$ changes by at most an arbitrarily chosen $\epsilon > 0$ for $t > t_2$.

4. Note that, for any value $\mu_{lo}(t_2) \in (0,1)$, the equation

$$\dot{x}(t) = G(w, x(t)) \tag{A.3}$$

with boundary condition $x(t_2) = \mu_{lo}(t_2)$ admits a solution that converges exponentially, as $t \to \infty$, to a positive quantity that can be written as $(-b + \sqrt{b^2 + c})$, where b and c are affine functions of w. The convergence is uniform in $\mu_{lo}(t_2)$.

5. Finally, using a comparison theorem (for instance, see Birkhoff and Rota (1969), page 25), $\mu_{lo}(t)$ is bounded by the solutions to (A.3) corresponding to w taking the highest and lowest values of $\mu_l(t)$ for $t > t_2$ (these are, of course, $\mu_l(t_2)$ and $\lim_{t\to\infty} \mu_h(t)$). By virtue of the previous step, for high enough t, these solutions are within $O(\epsilon)$ of the steady-state solution μ_{lo}.

□

Proof of Theorem 2: In order to calculate V_σ and P, we consider a particular agent and a particular time t, let τ_l denote the next (stopping) time at which that agent's intrinsic type changes, let τ_i denote the next (stopping) time at another investor with gain from trade is met, τ_m the next time a marketmaker is met, and let $\tau = \min\{\tau_l, \tau_i, \tau_m\}$. Then,

$$\begin{aligned}
V_{lo} &= E_t\Big[\int_t^{\tau} e^{-r(u-t)}(1-\delta)\,du + e^{-r(\tau_l-t)}V_{ho}1_{\{\tau_l=\tau\}} \\
&\qquad + e^{-r(\tau_i-t)}\,(V_{ln}+P)\,1_{\{\tau_i=\tau\}} \\
&\qquad + e^{-r(\tau_m-t)}\,(V_{ln}+B)\,1_{\{\tau_m=\tau\}}\Big] \\
V_{ln} &= E_t\left[e^{-r(\tau_l-t)}V_{hn}\right] \qquad\qquad (A.4)\\
V_{ho} &= E_t\left[\int_t^{\tau_l} e^{-r(u-t)}\,du + e^{-r(\tau_l-t)}V_{lo}\right] \\
V_{hn} &= E_t\Big[e^{-r(\tau_l-t)}V_{ln}1_{\{\tau_l=\tau\}} + e^{-r(\tau_i-t)}\,(V_{ho}-P)\,1_{\{\tau_i=\tau\}} \\
&\qquad + e^{-r(\tau_m-t)}\,(V_{ho}-A)\,1_{\{\tau_m=\tau\}}\Big],
\end{aligned}$$

where E_t denotes expectation conditional on the information available at time t. Differentiating both sides of Equation (A.4) with respect to t, we get (7).

In steady-state, $\dot{V}_\sigma = 0$ and hence (7) implies the following equations for the value functions and prices:

$$\begin{aligned}
V_{lo} &= \frac{(\lambda_u V_{ho} + 2\lambda\mu_{hn}P + \rho B + (2\lambda\mu_{hn}+\rho)V_{ln} + 1 - \delta)}{r + \lambda_u + 2\lambda\mu_{hn} + \rho} \\
V_{ln} &= \frac{\lambda_u V_{hn}}{r+\lambda_u} \qquad\qquad (A.5)\\
V_{ho} &= \frac{(\lambda_d V_{lo} + 1)}{r+\lambda_d} \\
V_{hn} &= \frac{(\lambda_d V_{ln} + (2\lambda\mu_{lo}+\rho)V_{ho} - 2\lambda\mu_{lo}P - \rho A)}{r+\lambda_d+2\lambda\mu_{lo}+\rho}
\end{aligned}$$

(We note that agents on the "long side" of market are rationed when they interact with the marketmaker, and, therefore, their trading intensity with the marketmaker is less than ρ. This does not affect (A.5), however, because the price is the reservation value.) Define $\Delta V_l = V_{lo} - V_{ln}$ and $\Delta V_h = V_{ho} - V_{hn}$ to be the reservation values. With this notation, the prices are determined by

$$\begin{aligned} P &= \Delta V_l(1-q) + \Delta V_h q \\ A &= \Delta V_h z + M(1-z) \\ B &= \Delta V_l z + M(1-z) \\ M &= \begin{cases} \Delta V_h & \text{if } s < \frac{\lambda_u}{\lambda_u+\lambda_d} \\ \Delta V_l & \text{if } s > \frac{\lambda_u}{\lambda_u+\lambda_d} \end{cases} \end{aligned} \tag{A.6}$$

and $M \in [\Delta V_l, \Delta V_h]$ if $s = \frac{\lambda_u}{\lambda_u+\lambda_d}$. Let

$$\begin{aligned} \psi_d &= \lambda_d + 2\lambda\mu_{lo}(1-q) + (1-\tilde{q})\rho(1-z) \\ \psi_u &= \lambda_u + 2\lambda\mu_{hn}q + \tilde{q}\rho(1-z)\ , \end{aligned}$$

where

$$\tilde{q} \begin{cases} = 1 & \text{if } s < \frac{\lambda_u}{\lambda_u+\lambda_d} \\ = 0 & \text{if } s > \frac{\lambda_u}{\lambda_u+\lambda_d} \\ \in [0,1] & \text{if } s = \frac{\lambda_u}{\lambda_u+\lambda_d}. \end{cases}$$

With this notation, we see that appropriate linear combinations of (A.5)–(A.6) yield

$$\begin{bmatrix} r+\psi_u & -\psi_u \\ -\psi_d & r+\psi_d \end{bmatrix} \begin{bmatrix} \Delta V_l \\ \Delta V_h \end{bmatrix} = \begin{bmatrix} 1-\delta \\ 1 \end{bmatrix}.$$

Consequently,

$$\begin{bmatrix} \Delta V_l \\ \Delta V_h \end{bmatrix} = \frac{1}{r}\begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{\delta}{r}\frac{1}{r+\psi_u+\psi_d}\begin{bmatrix} r+\psi_d \\ \psi_d \end{bmatrix}, \tag{A.7}$$

which leads to the price formula stated by the theorem.

Finally, we need to verify that any agent prefers, at any time, given all information, to play the proposed equilibrium trading strategy, assuming that other agents do. It is enough to show that an agent agrees to trade at the candidate equilibrium prices when contacted by an investor with whom there are potential gains from trade.

The Bellman principle for an agent of type lo in contact with an agent of type hn, is

$$\begin{aligned} P + V_{ln} \;\geq\; E_t\Bigg[& \int_t^{\tau} e^{-r(u-t)}(1-\delta)\,du + e^{-r(\tau_l - t)}V_{ho}1_{\{\tau_l=\tau\}} \\ & + e^{-r(\tau_i-t)}\left(V_{ln}+P\right)1_{\{\tau_i=\tau\}} \\ & + e^{-r(\tau_m-t)}\left(V_{ln}+B\right)1_{\{\tau_m=\tau\}}\Bigg], \end{aligned}$$

where $\tau = \min\{\tau_l, \tau_i, \tau_m\}$. This inequality follows from that fact that $\Delta V_h \geq P \geq \Delta V_l$. It says that selling the asset, consuming the price, and attaining the candidate value of a non-owner with low valuation, dominates (at least weakly) the value of keeping the asset, consuming its dividends and collecting the discounted expected candidate value achieved at the next time τ_m of a trading opportunity or at the next time τ_r of a change to a low discount rate, whichever comes first. There is a like Bellman inequality for an agent of type hn.

Now, to verify the sufficiency of the Bellman equations for individual optimality, consider any initial agent type $\sigma(0)$, any feasible trading strategy, θ, an

adapted process whose value is 1 whenever the agent owns the asset and 0 whenever the agent does not own the asset. The type process associated with trading strategy θ is denoted σ^θ. The cumulative consumption process C^θ associated with this trading strategy is given by

$$dC_t^\theta = \theta_t \left(1 - \delta 1_{\{\sigma^\theta(t)=lo\}}\right)\, dt - P\, d\theta_t. \tag{A.8}$$

Following the usual verification argument for stochastic-control, for any future meeting time τ^m, $m \in \mathbb{N}$, we have

$$V_{\sigma(0)} \geq E\left[\int_0^{\tau^m} e^{-rt}\, dC_t^\theta\right] + E\left[e^{-r\tau^m} V_{\sigma^\theta(\tau^m)}\right].$$

(This assumes without loss of generality that a potential trading contact does not occur at time 0.) Letting m go to ∞, we have $V_{\sigma(0)} \geq U(C^\theta)$. Because $V_{\sigma(0)} = U(C^*)$, where C^* is the consumption process associated with the candidate equilibrium strategy, we have shown optimality.

□

Proof of Theorem 3: The convergence of the masses μ to μ^* is easily seen using (A.1), whether λ or ρ tends to infinity. Let us concentrate on the prices.

1. If $s < \lambda_u/(\lambda_u + \lambda_d)$, then we see using (A.1) that $\lambda\mu_{hn}$ tends to infinity with λ, while $\lambda\mu_{lo}$ is bounded. Hence, Equation (A.7) shows that both ΔV_l and ΔV_h tend to r^{-1}, provided that $q > 0$. If $s > \lambda_u/(\lambda_u + \lambda_d)$, $\lambda\mu_{lo}$ tends to infinity with λ, while $\lambda\mu_{hn}$ is bounded. Hence, ΔV_l and ΔV_h tend to $r^{-1}(1-\delta)$, provided that $q < 1$. If $s = \lambda_u/(\lambda_u + \lambda_d)$, then $\lambda\mu_{hn} = \lambda\mu_{lo}$ tends to infinity with λ, and ΔV_l and ΔV_h tend to $r^{-1}(1 - \delta(1-q))$. In each case, the reservation values converge to the same value, which is a Walrasian price.

2. Equation (A.7) shows that both ΔV_l and ΔV_h tend to the Walrasian price $r^{-1}(1-\delta(1-\tilde{q}))$ as ρ approaches infinity.

3. When $z = 1$, $A^k - B^k$ increases with ρ because $A - B = \delta(r + \psi_u + \psi_d)^{-1}$ and both ψ_u and ψ_d decrease, since μ_{lo} and μ_{hn} do.

□

Proof of Theorem 4: Let the value function of a sophisticated type-σ investor be V_σ^s, and the value function of an unsophisticated type-σ investor be V_σ^u. These value functions and the prices (A^s, B^s, A^u, B^u) are computed as in (A.5)–(A.6), with the modification that the inter-dealer price M is different. For any fixed inter-dealer price M, an agent who meets the marketmaker with intensity ρ, and who sells as a lo type and buys as a hn type (i.e. with $\Delta V_l \leq M \leq \Delta V_h$) has value functions determined by

$$
\begin{aligned}
V_{ho}(r+\lambda_d) &= 1 + \lambda_d V_{lo} \\
V_{hn}(r+\lambda_d+\rho) &= \lambda_d V_{ln} + \rho(V_{ho} - [z\Delta V_h + (1-z)M]) \\
V_{ln}(r+\lambda_u) &= \lambda_u V_{hn} \\
V_{lo}(r+\lambda_u+\rho) &= 1 - \delta + \lambda_u V_{ho} + \rho(V_{ln} + [z\Delta V_l + (1-z)M]).
\end{aligned}
$$

The system reduces to

$$
\begin{aligned}
\Delta V_h(r+\lambda_d+\rho(1-z)) &= 1 + \lambda_d \Delta V_l + \rho(1-z)M \\
\Delta V_l(r+\lambda_u+\rho(1-z)) &= 1 - \delta + \lambda_u \Delta V_h + \rho(1-z)M,
\end{aligned}
$$

which implies that

$$\begin{bmatrix} \Delta V_l \\ \Delta V_h \end{bmatrix} = \frac{1+\rho(1-z)M}{r+\rho(1-z)} \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{\delta}{r+\rho(1-z)} \frac{1}{r+\lambda_u+\lambda_d+\rho(1-z)} \begin{bmatrix} r+\lambda_d+\rho(1-z) \\ \lambda_d \end{bmatrix}. \tag{A.9}$$

Hence, this agent faces a bid-ask spread of

$$z(\Delta V_h - \Delta V_l) = \frac{z\delta}{r+\lambda_u+\lambda_d+\rho(1-z)}.$$

We show below, for each case, that M is given by

$$M = \begin{cases} \Delta V_h^s & \text{if } s < \mu^s \frac{\lambda_u}{\lambda_u+\lambda_d} \\ \Delta V_h^u & \text{if } \mu^s \frac{\lambda_u}{\lambda_u+\lambda_d} < s < \frac{\lambda_u}{\lambda_u+\lambda_d} \\ \Delta V_l^u & \text{if } \frac{\lambda_u}{\lambda_u+\lambda_d} < s < 1-\mu^s \frac{\lambda_d}{\lambda_u+\lambda_d} \\ \Delta V_l^s & \text{if } 1-\mu^s \frac{\lambda_d}{\lambda_u+\lambda_d} < s. \end{cases} \tag{A.10}$$

Case (a). Consider first the case $s < \mu^s \lambda_u/(\lambda_u+\lambda_d)$. The claim is that it is an equilibrium that the unsophisticated investors do not own any shares and do not trade. Assuming this to be true, the market has only sophisticated investors, the interdealer price is $M = \Delta V_h^s$, and the buyers are rationed.

It remains to be shown that, with this interdealer price, there is no price at which marketmakers will sell and unsophisticated investors will buy. First of all, we note that the optimal response of an investor to the Markov (time-independent) investment problem can be chosen to be Markov, which means that one only needs to check the payoffs from Markov strategies that stipulate the same probability of trade for a give type at any time. The linearity of the problem further allows one

to assume that the trading probability is 1 or zero. (When indifferent, the choice does not matter, so we may assume a corner solution.)

There are three possible Markov strategies for the unsophisticated investor that involve buying: buying as type h and selling as type l, buying as type l and selling as type h, and buying and holding (never selling).

If the unsophisticated investor buys as an h type and sells as an l type, then her value function satisfies (A.9), implying that $\Delta V_h^u < \Delta V_h^s = M$ since $\rho^u < \rho^s$. The reservation values are even lower if she buys as an l and sells as an h type. Finally, if the unsophisticated investor buys and never sells, then her value function is also smaller than M. This is inconsistent with trading with the marketmaker, meaning that she never buys.

Case (b). For the case $\mu_h^s < s < \mu_h$, the equilibrium is given by an inter-dealer price of $A^u = M = \Delta V_h^u = A(\rho^u)$. This is also the price at which unsophisticated hn-agents buy from the marketmaker, and these agents are rationed. The sophisticated types hold a total $\mu_h^s = \mu^s \lambda_u/(\lambda_u + \lambda_d)$ of the supply, while the unsophisticated types hold the rest. This is clearly an equilibrium for the unsophisticated types. We have to ensure that sophisticated t ypes also behave optimally. In particular, we have to check that $\Delta V_l^s \leq M \leq \Delta V_h^s$. For this, we use (A.7) and (A.9):

$$\Delta V_l^s \leq M \Leftrightarrow$$

$$\frac{1 + \rho^s(1-z)M}{r + \rho^s(1-z)} - \frac{\delta(r + \lambda_d + \rho^s(1-z))}{r + \rho^s(1-z)} \frac{1}{r + \lambda_u + \lambda_d + \rho^s(1-z)} \leq M \Leftrightarrow$$

$$\frac{r + \lambda_d + \rho^s(1-z)}{r + \lambda_u + \lambda_d + \rho^s(1-z)} \geq \frac{\lambda_d}{r + \lambda_u + \lambda_d + \rho^u(1-z)}$$

where the last inequality is satisfied because $\rho^s \geq \rho^u$. Similarly, it can be verified that $M \leq \Delta V_h^s$ using the same formulae.

Case (c). The remaining two cases are dual to the ones that we just proved. To see this, take the following new perspective of an agent's problem: An agent considers "acquiring" non-ownership (that is, selling). The number of "shares" of non-ownership is $1 - s$. If an l-type acquires non-ownership then he gets a "dividend" of $-(1 - \delta)$ (that is, he gives up a dividend of $1 - \delta$). If a h-type acquires non-ownership then he gets a "dividend" of -1. Said differently, he gets a dividend of $-(1-\delta)$ like the l-type, and, additionally, he has a cost of δ. Hence, from this perspective h and l types are reversed, and the supply of "shares" is $1 - s$.

This explains why the equilibria in the latter two cases are the mirror images of the equilibria in the former two cases. In particular, if $\frac{\lambda_u}{\lambda_u+\lambda_d} < s < 1 - \mu^s \frac{\lambda_d}{\lambda_u+\lambda_d}$, then both sophisticated and unsophisticated investors trade, and the unsophisticated l type is rationed.

If $1 - \mu^s \frac{\lambda_d}{\lambda_u+\lambda_d} < s$, each unsophisticated investor owns a share and does not trade. (Using the alternative perspective, they are out of the market for non-ownership). The sophisticated investors hold the remaining $(1 - \mu^s)$ shares, they trade, and the selling sophisticated investors are rationed.

□

Proof of Theorem 5:

There exists a number ρ^M that maximizes (16) since π^M is continuous and $\pi^M(\rho) \to -\infty$ as $\rho \to \infty$.

We are looking for $\rho^C \geq 0$ such that

$$\Gamma'(\rho^C) = rE \int_0^\infty \mu_m(\rho^C)(A(\rho^C) - B(\rho^C))e^{-rt}\, dt. \tag{A.11}$$

Consider how both the left and right-hand sides depend on ρ. The left-hand side

is 0 for $\rho = 0$, increasing, and tends to infinity as ρ tends to infinity. For $z = 0$, $A(t, \rho) - B(t, \rho) = 0$ everywhere, so the the right-hand side (RHS) is zero, and, therefore, the unique solution to (A.11) is clearly $\rho^C = 0$. For $z > 1$, the RHS is strictly positive for $\rho = 0$. Further, the steady-state value of the RHS can be seen to be decreasing, using the fact that μ_m is decreasing in ρ, and using the explicit expression for the spread provided by (A.7). Further, by continuity and still using (A.7), there is $\varepsilon > 0$ and T such that $\frac{\partial}{\partial\rho}\mu_m(A - B) < -\varepsilon$ for all $t > T$ and all r. Further, note that $r\exp(-rt)$ is a density function for all r, and that the closer r is to zero, the more weight is given to high values of t (that is, the more important is the steady-state value for the integral). Therefore, the RHS is also decreasing in ρ for any initial condition on μ if r is small enough. These results yield the existence of a unique solution.

To see that $\rho^C > \rho^M$ when $z = 1$, consider the first-order conditions that determine ρ^M:

$$\begin{aligned}\Gamma'(\rho^M) \;=\; rE\int_0^\infty \Big[&\mu_m(t,\rho^M)(A(t,\rho^M) - B(t,\rho^M)) \\ &+\rho^M \frac{\partial}{\partial \rho^M}\left(\mu_m(t,\rho^M)(A(t,\rho^M) - B(t,\rho^M))\right)\Big] e^{-rt}\, dt.\end{aligned} \tag{A.12}$$

The integral of the first integrand term on the right-hand side of (A.12) is the same as that of (A.11), and that of the second is negative for small r. Hence, the right-hand side of (A.12) is smaller than the right-hand side of (A.11), implying that $\rho^C(1) > \rho^M$.

To see that $\rho^C(z)$ is increasing in z, we use the Implicit Function Theorem and the dominated convergence theorem to compute the derivative of $\rho^C(z)$ with

respect to z, as

$$\frac{rE\int_0^\infty \mu_m(\rho^C)(A_z(\rho^C,z)-B_z(\rho^C,z))e^{-rt}\,dt}{\Gamma''(\rho^C)-rE\int_0^\infty \frac{d}{d\rho}\mu_m(\rho^C)(A(\rho^C,z)-B(\rho^C,z))e^{-rt}\,dt}. \tag{A.13}$$

If we use the steady-state expressions for μ, A, and B, this expression is seen to be positive because both the denominator and the numerator are positive. Hence, it is also positive with any initial masses if we choose r small enough.

□

Proof of Theorem 6: *(i)* The first part of the theorem, that the monopolistic marketmaker's search intensity does not affect investors when they can't search for each other, follows from (A.5), which shows that investor's utility is independent of ρ.

(ii) We want to prove that the investor welfare is decreasing in ρ, which directly implies that the marketmaker over-invests in intermediation services.

We introduce the notation $\Delta V_o = V_{ho} - V_{lo}$, $\Delta V_n = V_{hn} - V_{ln}$, and $\phi = \Delta V_h - \Delta V_l = \Delta V_o - \Delta V_n$, and start by proving a few general facts about the marketmaker spread, ϕ.

The dynamics of ϕ are given by the ordinary differential equation (ODE)

$$\dot{\phi}_t = (r+\lambda_d+\lambda_u+2\lambda(1-q)\mu_{lo}+2\lambda q\mu_{hn})\phi_t - \delta,$$

Let $R = r+\lambda_d+\lambda_u+2\lambda(1-q)\mu_{lo}+2\lambda q\mu_{hn}$. The equation above readily implies that

$$\frac{\partial\dot{\phi}_t}{\partial\rho} = R\frac{\partial\phi_t}{\partial\rho} + \left(2\lambda(1-q)\frac{\partial\mu_{lo}(t)}{\partial\rho} + 2\lambda q\frac{\partial\mu_{hn}(t)}{\partial\rho}\right)\phi_t. \tag{A.14}$$

This can be viewed an ODE in the function $\frac{\partial\phi}{\partial\rho}$ by treating ϕ_t is a fixed function.

It can be verified that $0 < \frac{\partial \phi}{\partial \rho} < \infty$ in the limit as $t \to \infty$, that is, in steady state. Further, a simple comparison argument yields that $\frac{\partial \mu_{lo}(t)}{\partial \rho} = \frac{\partial \mu_{hn}(t)}{\partial \rho} < 0$. Hence, the solution to the linear ODE (A.14) is positive since

$$\frac{\partial \phi_t}{\partial \rho} = -\int_t^\infty e^{-R(u-t)} \left(2\lambda(1-q) \frac{\partial \mu_{lo}(u)}{\partial \rho} + 2\lambda q \frac{\partial \mu_{hn}(u)}{\partial \rho} \right) \phi_u \, du > 0.$$

Consider now the case $q = 1$, for which, since $V_{hn} = V_{ln} = 0$,

$$\dot{V}_{ho}(t) = rV_{ho}(t) + \lambda_d \phi_t - 1 \,.$$

Differentiating both sides with respect to ρ and using arguments as above, we see that $\frac{\partial V_{ho}(t)}{\partial \rho} < 0$ since $\frac{\partial \phi_t}{\partial \rho} > 0$. Consequently, $V_{lo}(t) = V_{ho}(t) - \phi_t$ also decreases in ρ.

If $q = 0$, then (A.5) shows that V_{lo} and V_{ho} are independent of ρ. Further,

$$\dot{V}_{ln}(t) = rV_{ln}(t) + \lambda_u(\phi_t - \Delta V_o(t)).$$

As above, we differentiate with respect to ρ and conclude that $V_{ln}(t)$ decreases in ρ since $\frac{\partial \phi_t}{\partial \rho} > 0$ and $\Delta V_o(t)$ is independent of ρ. Consequently, $V_{hn}(t) = V_{ln}(t) - \phi_t + \Delta V_o(t)$ also decreases in ρ.

□

Proof of Theorem 7:

To see that $z^I > 0$, we note that with $\rho = \rho^C(z)$,

$$\frac{d}{dz} w^I \mid_{z=0} = -\delta E \int_0^\infty \frac{d}{d\rho} \mu_{lo}(t, \rho) e^{-rt} \, dt \, \frac{d\rho^C}{dz} > 0,$$

where we have used that $\rho^C(0) = 0$, that $\frac{d\rho^C}{dz} > 0$ at $z = 0$ (see (A.13)), that

$A - B = 0$ if $z = 0$, and that for all t, $\frac{d}{d\rho}\mu_{lo}(t, \rho) < 0$.

To prove that $z^I < z^S \leq z^M = 1$, it suffices to show that the marketmaker welfare is increasing in z, which follows from

$$\begin{aligned}
\frac{d}{dz} w^M &= \rho \frac{d}{dz} \left[E \int_0^\infty \mu_{lo}(a-b) e^{-rt}\, dt \right] \\
&= \frac{\rho}{r} \frac{d}{dz} \Gamma'(\rho^C(z)) \\
&= \frac{\rho}{r} \Gamma''(\rho^C(z)) \frac{d\rho^C}{dz} > 0,
\end{aligned}$$

suppressing the arguments t and ρ from the notation, where we have used twice the fact that $\Gamma'(\rho) = rE \int_0^\infty \mu_{lo}(A-B)e^{-rt}\, dt$ if $\rho = \rho^C(z)$, and that $\frac{d\rho^C}{dz} > 0$ (Theorem 5).

□

References

Amihud, Y. and H. Mendelson (1980). Dealership Markets: Market Making with Inventory. *Journal of Financial Economics 8*, 31–53.

Amihud, Y. and H. Mendelson (1986). Asset Pricing and the Bid-Ask Spread. *Journal of Financial Economics 17*, 223–249.

Bagehot, W. (1971). The Only Game in Town. *The Financial Analysts Journal 27*, 12–14.

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. *Journal of Finance 54*, 1–34.

Bester, H. (1988). Bargaining, Search Costs and Equilibrium Price Distributions. *Review of Economic Studies 55*, 201–214.

Bester, H. (1989). Noncooperative Bargaining and Spatial Competition. *Econometrica 57*, 97–113.

Bhattacharya, S. and K. M. Hagerty (1987). Dealerships, Trading Externalities, and General Equilibrium. In E. Prescott and N. Wallace (Eds.), *Contractual Arrangements for Intertemporal Trade*, Chapter Minnesota Studies in Macroeconomics Series, vol. 1, pp. 81–104. Minneapolis: University of Minnesota Press.

Binmore, K. G. and M. J. Herrero (1988). Matching and Bargaining in Dynamic Markets. *Review of Economic Studies 55*, 17–31.

Birkhoff, G. and G.-C. Rota (1969). *Ordinary Differential Equations*. John Wiley & Sons.

Constantinides, G. M. (1986). Capital Market Equilibrium with Transaction Costs. *Journal of Political Economy 94*, 842–862.

Dai, Q. and K. Rydqvist (2003). How do Buyers and Sellers Divide The Surplus. Unpublished working paper. Binghamton University.

DeMarzo, P., M. Fishman, and K. Hagerty (2000). The Enforcement Policy of a Self-Regulatory Organization. Unpublished working paper. Graduate School of Business, Stanford University.

Diamond, P. (1982). Aggregate Demand Management in Search Equilibrium. *Journal of Political Economy 90*, 881–894.

Duffie, D., N. Gârleanu, and L. H. Pedersen (2003). Valuation in Other-the-Counter Markets. Graduate School of Business, Stanford University.

Duffie, D. and Y. Sun (2004). The Exact Law of Large Numbers for Pairwise Random Matching. Unpublished working paper. Graduate School of Business, Stanford University.

Gale, D. (1986a). Bargaining and Competition Part I: Characterization. *Econometrica 54*, 785–806.

Gale, D. (1986b). Bargaining and Competition Part II: Existence. *Econometrica 54*, 807–818.

Gale, D. (1987). Limit Theorems for Markets with Sequential Bargaining. *Journal of Economic Theory 43*, 20–54.

Garman, M. (1976). Market Microstructure. *Journal of Financial Economics 3*, 257–275.

Gehrig, T. (1993). Intermediation in Search Markets. *Journal of Economics and Management Strategy 2*, 97–120.

Glosten, L. R. and P. R. Milgrom (1985). Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. *Journal of Financial Economics 14*, 71–100.

Harris, M. (1979). Expectations and Money in a Dynamic Exchange Model. *Econometrica 47*, 1403–1419.

Ho, T. and H. R. Stoll (1981). Optimal Dealer Pricing Under Transactions and Return Uncertainty. *Journal of Financial Economics 9*, 47–73.

Hosios, A. J. (1990). On the Efficiency of Matching and Related Models of Search and Unemployment. *Review of Economic Studies 57*, 279–298.

Huang, R. D. and H. R. Stoll (1996). Dealer versus Auction Markets: A Paired Comparison of Execution Costs on NASDAQ and the NYSE. *Journal of*

Financial Economics 41 (3), 313–357.

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

Kreps, D. (1990). *A Course in Microeconomic Theory*. Princeton University Press.

Kyle, A. S. (1985). Continuous Auctions and Insider Trading. *Econometrica 6*, 1315–1335.

Lamoureux, C. G. and C. R. Schnitzlein (1997). When It's Not the Only Game in Town: The Effect of Bilateral Search on the Quality of a Dealer Market. *Journal of Finance 52*, 683–712.

McLennan, A. and H. Sonnenschein (1991). Sequential Bargaining as a Noncooperative Foundation for Walrasian Equilibrium. *Econometrica 59*, 1395–1424.

Moen, E. R. (1997). Competitive Search Equilibrium. *Journal of Political Economy 105*, 385–411.

Moresi, S. (1991). Three Essays in Economic Theory. Ph.D. Thesis, MIT.

Mortensen, D. T. (1982). Property Rights and Efficiency in Mating, Racing, and Related Games. *American Economic Review 72*, 968–979.

Nash, J. (1950). The Bargaining Problem. *Econometrica 18*, 155–162.

New York Stock Exchange (2001). Fact Book 2001.

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

Protter, P. (1990). *Stochastic Integration and Differential Equations*. New York: Springer-Verlag.

Rubinstein, A. and A. Wolinsky (1985). Equilibrium in a Market with Sequential Bargaining. *Econometrica 53*, 1133–1150.

Rubinstein, A. and A. Wolinsky (1987). Middlemen. *Quarterly Journal of Economics 102*, 581–594.

Sun, Y. (2000). On the Sample Measurability Problem in Modeling Individual Risks. Department of Mathematics and Center for Financial Engineering, National University of Singapore.

Trejos, A. and R. Wright (1995). Search, Bargaining, Money, and Prices. *Journal of Political Economy 103*, 118–140.

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

Vayanos, D. and T. Wang (2002). Search and Endogenous Concentration of Liquidity in Asset Markets. Working Paper, MIT.

Weill, P.-O. (2002). The Liquidity Premium in a Dynamic Bargaining Market. Working Paper, Stanford University.

Yavaş, A. (1996). Search and Trading in Intermediated Markets. *Journal of Economics and Management Strategy 5*, 195–216.