NBER WORKING PAPER SERIES

THE SEARCH THEORY OF OTC MARKETS

Pierre-Olivier Weill

Working Paper 27354
http://www.nber.org/papers/w27354

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2020

Prepared for the Annual Review of Economics. This paper benefited from comments and suggestions by Stephen Morris (the Editor), David Lindsay, Semih Uslu, Mengbo Zhang, and Diego Zuniga. David Lindsay provided expert research assistance. I would like to thank Batchimeg Sambalaibat for sharing data about trading volume across markets. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

The search theory of OTC markets
Pierre-Olivier Weill
NBER Working Paper No. 27354
June 2020
JEL No. G0,G12

ABSTRACT

I review the recent literature that applies search-and-matching theory to the study of Over-the-Counter (OTC) financial markets. I formulate and solve a simple model in order to illustrate the typical assumptions and economic forces at play in existing work. I then offer thematic tours of the literature and, in the process, discuss avenues for future research.

Pierre-Olivier Weill
Department of Economics
University of California, Los Angeles
Bunche Hall 8283
Los Angeles, CA 90095
and NBER
poweill@econ.ucla.edu

1 Introduction

Centralized financial markets are typically organized as limit order books: all-to-all trading platforms with executable quotes. In contrast, over-the-counter (OTC) markets are harder to categorize. To use Lucas (1989, page 272) famous paraphrase of Tolstoy, "all centralized markets are the same, but each OTC market is unique in its own way." However, OTC markets do have common features. In particular, investors trade in small groups and not in all-to-all auctions. For example, they trade with only one dealer or in a Request For Quote (RFQ) auction with a few dealers. There is no pre-trade transparency: quotes, if any, are not executable and so in principle are up for negotiation. Post-trade transparency may also be limited in the sense that investors may have to negotiate terms of trades under imperfect and often asymmetric information about overall market conditions.

Examples of assets that trade mostly in OTC markets include fixed-income securities, various types of derivatives, repos and federal funds loans (SIFMA, 2018; ISDA, 2018). Even for assets that mostly trade in centralized markets, OTC markets may attract an economically significant fraction of volume. For example, traditionally, large "blocks" of equity have traded bilaterally in the upstairs market (see Harris, 2003, chapter 15). Also, in the last decade, the equity market has become extremely fragmented across multiple trading venues, some centralized and some OTC (Tuttle, 2014; Hatheway, Kwan, and Zheng, 2017). Figure 1 breaks down the outstanding supply of publicly traded domestic securities between centralized (black bar) and OTC markets (grey bar), in 2018.[1] It reveals that the value of all domestic assets traded in OTC markets is very large, over 50,000 billion USD. It exceeds the value of assets traded in centralized markets, by more than 20,000 billion USD.

Some have argued that OTC markets are doomed to disappear: ever improving trading technologies will inevitably make markets more centralized. But the historical record suggests otherwise. Biais and Green (2019) show that, prior to the 1940s, corporate and municipal bonds traded mostly in centralized markets, but that the trading volume migrated over time towards OTC markets. Riggs, Onur, Reiffen, and Zhu (2018) document that, in the market for credit default swaps, investors can now freely choose between different types of trading mechanisms, some more centralized than others. They find that the most centralized mechanism, a limit-order book, attracts very little volume.

[1]For equities, I use volume share to calculate the fraction of supply traded in centralized and in OTC markets. For Treasuries, I use estimates of the volume share of electronic and voice trading, for on-the-run and off-the-run securities. See online appendix for underlying assumptions and details.

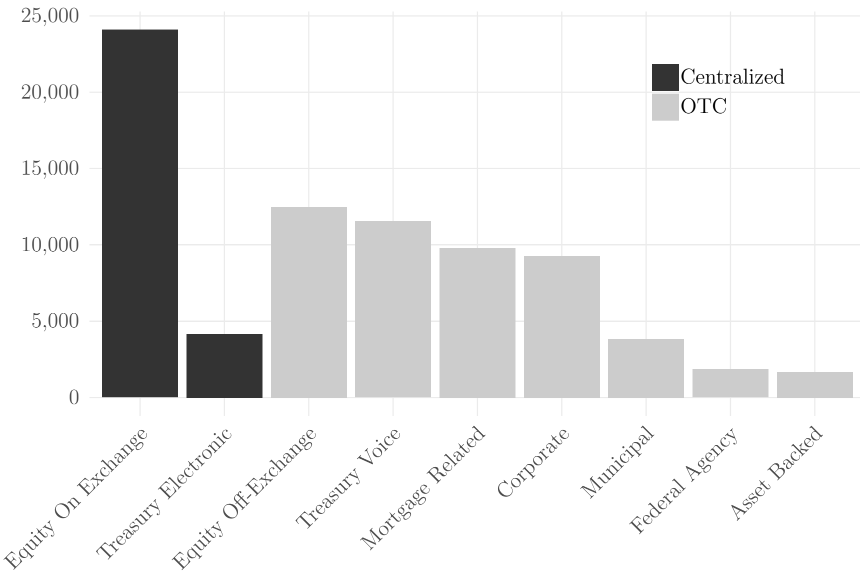

Figure 1: Security supply outstanding in 2018, broken down between Centralized and OTC markets, Billions of 2019 USD.

Why are we interested in the economics of OTC markets? Over the last two decades, OTC markets have been at the center of several policy and regulatory debates, all of which underscore the need for rigorous theoretical analyses. For example, in the early 2000s, reporting systems were put in place in the OTC markets for corporate and municipal bonds, with the goal of disseminating price information to the market. This led to an important debate on the economic value of post-trade transparency (see Bessembinder and Maxwell, 2008). Several OTC markets were at the core of the 2008 financial crisis, which ignited vigorous policy debates. For example, how should policymakers and regulators respond to the collapse of trading volume in several asset classes traded in OTC markets? Should the government engage in ex-post intervention such as large-scale purchase of Mortgage Based Securities (MBS)? Should there be a mandate to implement more centralized and transparent trading mechanisms? Another debate concerned the appropriate policy response to the build up of systemic risk in the OTC market for credit derivatives. In particular, should these derivatives be centrally cleared, so as to concentrate counterparty credit risk in central clearing parties?

OTC markets also matter for broader macroeconomic questions. For example, given that most fixed-income securities are traded OTC, these markets impact the cost of capital of firms and sovereigns. For the most part, conventional and unconventional monetary policy is implemented in OTC markets for Treasuries, Repo, Federal Funds loans, and Mortgage related securities. Finally, a commonly held view is that OTC markets have amplified and propagated negative shocks during the financial crisis of 2008.

Where does the theory of OTC market stand? Most theoretical analyses of OTC markets rely on tools from either search or network theory, and sometimes both at the same time. Two elementary observations motivate these analyses. First, trade is fragmented within small groups of investors: for example, trade may occur bilaterally between two dealers in a voice transaction, or multilaterally between one customer and a small number of dealers in a RFQ auction. Second, an investor is not forced to trade with any group in the OTC marketplace. Instead, an investor can choose with whom and when to trade: for example, a customer can choose between several dealers at any point in time, and can choose to trade later with some other dealer. This blurs the distinction between centralized and decentralized markets: the ability to choose with whom and when to trade effectively brings investors together, even in the absence of a centralized all-to-all marketplace.[2]

Search theory provides natural and tractable frameworks to study this choice problem in equilibrium, in particular to derive implications for asset prices, transaction costs, trading volume, as well as the supply of intermediation services. Indeed, in a search model, investors trade in small groups, often bilaterally, and they have explicit options to delay trade and find some other counterparty in the market. Since OTC markets are sometimes dominated by a few dealers, network theory is appealing too for more detailed and precise analyses of strategic interactions amongst a finite number of investors. One advantage of search over network theory is tractability, as it facilitates the analysis of dynamic trading, pricing, and of the market response to aggregate shocks. As is often the case, though, it is not wise to be dogmatic about modeling frameworks: whether a search or a network model is more appropriate ultimately depends on the question at hand, and on the economic insights generated. In this review article, I will rely on my own expertise and focus most of my attention on search-theoretic models.[3]

[2]In fact, one can construct theoretical examples in which a decentralized marketplace achieves the same or nearly the same outcome as a centralized one. See, for example, the vanishing frictions limits in the dynamic search model of Duffie, Gârleanu, and Pedersen (2005), or the static network-theoretic model of Malamud and Rostek (2017), or the hybrid model of Atkeson, Eisfeldt, and Weill (2015).

[3]Models of OTC markets building on the financial network literature have been proposed by Babus and Hu (2018); Babus and Kondor (2018); Farboodi (2014); Gofman (2014); Gofman (2017); Wang (2018); Malamud and Rostek (2017); Babus and Parlatore (2018); Babus (2019); Babus and Farboodi (2018); Manea (2018); Eisfeldt, Herskovic, Rajan, and Siriwardane (2018), Aymanns, Co-Pierre, and Golub (2018). Some work also use an hybrid approach, blending element of search and network models, for example Atkeson, Eisfeldt, and

When applying search theory to financial markets, some scholars experience an uncomfortable stretch of their imagination. They argue that, in practice, investors cannot possibly face an economically meaningful search problem because they either know of their potential counterparties or can quickly learn about them. Several arguments may alleviate this concern. First, even if investors know the full directory of their counterparties, they do not know their trading needs: in particular, finding a counterparty willing to trade large quantities or customized derivative products may take a long time. Second, what makes the search problem economically significant is not the physical search time per se, but the opportunity cost of delaying trade. The magnitude of the price concession that financially distressed institutions are willing to concede (Duffie, 2010) suggests that these costs can be quite large. Third, an insight from theory is that the economic significance of search not only depends on the physical time to find a counterparty, but also on the pricing mechanism. It is well known from the Diamond Paradox (Diamond, 1971) that even small time delays can have a very large impact on prices when investors have low bargaining power. Thus, what ultimately matters for valuation is a bargaining-adjusted measure of search times. Fourth, while a search market is clearly an abstraction from the true underlying trade and price formation mechanism, so is a Walrasian market. The key advantage of the search market is to explain the equilibrium determination of economic phenomena ignored in Walrasian models, such as trading delays, transaction costs, price dispersion, or mismatch. Search models are not inconsistent with Walrasian models: in fact, a desirable property of many search models is to deliver Walrasian prices as frictions vanish.

The remainder of this paper has two parts. In Section 2 I first analyze a benchmark search model of OTC market. In Section 3, I use this model as a basis to review the existing literature and discuss avenues for future research.

2 A benchmark model

In this section, I formulate and solve a simple benchmark model to illustrate some of the typical assumptions and economic forces discussed in the literature following the pioneering work of Duffie, Gârleanu, and Pedersen (2005). The setup builds on Hugonnier, Lester, and Weill (2014), but with a simpler matching and price-setting mechanism.

Weill (2015), Colliard and Demange (2014); Colliard, Foucault, and Hoffmann (2018); Chang and Zhang (2018), and Dugast, Üslü, and Weill (2019).

2.1 The setup

Assets and agents. Time is continuous and runs forever. There is one asset in positive supply $s \in (0, 1)$. The economy is populated by two types of risk-neutral agents who discount the future at rate $r > 0$: a measure μ_d of dealers, and a measure μ_c of customers normalized to $\mu_c = 1$. Dealers cannot hold any inventory. A customer can hold either zero or one unit of the asset, in which case she derives a utility flow δ. To create a demand for trading assets, I assume that customers' utility flows vary independently over time: namely, at independent Poisson arrival times with intensity γ, a customer with current utility flow δ draws a new utility flow δ' in a compact interval according to the strictly positive probability density $f(\delta')$.

The assumption that customers derive some time-varying and idiosyncratic utility flow for assets has become standard in the literature, but it may not be obvious to interpret in the context of financial markets. The literature has provided several natural interpretations and micro-foundations. First, heterogenous utility flows can reflect agents' heterogenous beliefs about the distribution of asset future payoffs, as in Duffie, Gârleanu, and Pedersen (2002) or Hugonnier (2012). Second, heterogenous utility flows may arise from agents' heterogenous hedging needs. Imagine for example that agents are risk averse and hold some outside asset that is correlated with the asset under study – e..g, a corporate bond dealer who holds a large amount of Treasury bonds. Then large holdings of the outside asset depress the marginal value for the asset under study, generating a small δ, and vice versa (see the empirical study of Newman and Rierson, 2003). Of course, making this argument precise requires additional work, provided for example by Duffie, Gârleanu, and Pedersen (2007) and Vayanos and Weill (2008), and especially by the dynamic programming arguments of Praz (2015). Finally, in recent monetary models, customers have heterogeneous consumption opportunities which they can finance by selling assets in OTC markets (see, among others, Geromichalos and Herrenbrueck, 2016). The heterogenous consumption opportunities translate into heterogenous indirect utilities for assets.

Market. As first proposed by Duffie, Gârleanu, and Pedersen (2005), I will assume that the OTC market is semi-centralized: competitive dealers can trade instantly in a centralized inter-dealer market, but customers must search for dealers in order to trade. Specifically, customers randomly contact dealers at independent Poisson arrival times with intensity λ, and bargain over the terms of trade in a manner specified below.

Semi-centralized markets are realistic: in practice, customers must often trade through intermediaries, and much of the trading volume is made up of either customer-to-dealer trades, or dealer-to-dealer trades (see for example Atkeson, Eisfeldt, and Weill, 2013, for credit derivatives). Moreover, inter-dealer trades may be viewed as more competitive, as they are empirically associated with smaller price dispersion (Li and Schürhoff, 2019). The semi-centralized market assumption has also proved to be remarkably tractable in extensions of the basic model.

2.2 Bargaining between customers and dealers

Let $V_q(\delta)$ denote the maximum attainable utility of a customer who holds $q \in \{0,1\}$ units of the asset, with current utility flow δ. Consider a type-δ customer who owns an asset, $q = 1$, and contacts a dealer. Given the $q \in \{0,1\}$ restriction on asset holdings, the only possible trade is that the customer sells her asset to the dealer. If the trade occurs, the customer receives utility $B(\delta) + V_0(\delta)$, for some bid price $B(\delta)$ to be determined. If the trade does not occur, the customer continues her search and so receives the utility $V_1(\delta)$. Hence, the net utility of the customer is:

$$B(\delta) + V_0(\delta) - V_1(\delta) = B(\delta) - \Delta V(\delta),$$

where $\Delta V(\delta) \equiv V_1(\delta) - V_0(\delta)$ denotes the reservation value of the customer. The net utility of the dealer is, on the other hand:

$$P - B(\delta),$$

where P denotes the inter-dealer price, to be determined later in equilibrium. Following the literature, I determine the bid price via generalized Nash Bargaining:

$$B(\delta) = \arg\max \left(B - \Delta V(\delta)\right)^{1-\theta} \left(P - B\right)^{\theta}$$

with respect to B, subject to $B - \Delta V(\delta) \geq 0$, $P - B \geq 0$, and given some $\theta \in [0,1]$ representing the dealer's bargaining power. One sees that the constraint set is non-empty if and only if $P - \Delta V(\delta) \geq 0$. That is, there are gains from trade between the customer and the dealer if and only if the inter-dealer price is larger than the customer's reservation value. When the constraint set is non-empty, the first-order necessary and sufficient conditions yield:

$$B(\delta) = \theta \Delta V(\delta) + (1-\theta)P. \tag{1}$$

Hence, the bid price lies between the customer's reservation value and the inter-dealer price. If the dealer has strong bargaining power, $\theta \simeq 1$, then the bid price is close to the customer's reservation value, that is, the price set by a fully discriminating monopsonist. Vice versa, if the customer has strong bargaining power, $\theta \simeq 0$, then the price is competitive.

Proceeding the same way for a customer who does not own an asset, we obtain that there are gains from trade if and only if the customer's reservation value is larger than the inter-dealer price, $\Delta V(\delta) \geq P$. Moreover, the ask price is given by the same formula as the bid price:

$$A(\delta) = \theta \Delta V(\delta) + (1 - \theta)P. \tag{2}$$

Price setting mechanisms. The bid (1) and the ask (2), can be viewed as the outcomes of realistic strategic bargaining protocols and price setting mechanisms in OTC market.

For example, as is well known, Nash bargaining prices arise when the customer and the dealer engage in a fully specified dynamic bargaining game with alternative offers, as in Rubinstein (1982). One important difference is that the bargaining weight, θ, is now endogenous and depends on other model parameters.

The Nash bargaining payoffs are also obtained if we assume that each customer simultaneously contacts several dealers to receive price quotes, in a game à-la Burdett and Judd (1983). This is appealing for at least two reasons. First, since customers can choose between several standing offers, this price-setting mechanism captures in a simple way the possibility of recall. Second, this mechanism also closely resembles an RFQ auction, which are now common in OTC markets: for example, in the corporate bond or credit default swap markets, customers can request quotes from several dealers at the same time (Fermanian, Guéant, and Pu, 2016; Riggs, Onur, Reiffen, and Zhu, 2018). To be more specific, let us consider the following stylized representation of an RFQ auction. Assume that each customer sends $n \geq 2$ simultaneous quote requests to randomly chosen dealers, but dealers independently decline the request with probability $\pi \in (0, 1)$ for exogenous reasons, perhaps because they are busy with other clients (for example Riggs, Onur, Reiffen, and Zhu show that, in their CDS data, dealers do not respond to RFQ about 10 percent of the times). Dealers are uncertain about the number of offers received by customers. As a result, the arguments of Burdett and Judd (see online appendix) shows that they play a mixed-strategy, randomizing their price quotes from a non-atomic distribution. Although the resulting price dispersion is different from what would be obtained with Nash bargaining, the average payoffs are the same. The average transaction price is $\theta \Delta V(\delta) + (1 - \theta)P$, where θ is the customer's probability of receiving just one quote, conditional on receiving some quote. Thus, under this stylized model of RFQ, most predictions of the bilateral trading model remain unchanged. See Lester, Shourideh, Venkateswaran, and Zetlin-Jones (2018) for Burdett and Judd pricing under asymmetric information, Duffie, Dworczak, and Zhu (2017) and Glebkin et al. (2019) for closely related price setting mechanism.

The price setting protocols described above are opaque, in the sense that customers must search for dealers in order to discover or bargain over prices. However, in several OTC markets, regulators have promoted price transparency. A convenient way to model a transparent price-setting protocol is to assume that customers can direct their search towards the price posted by dealers, as in models of competitive search. Lester, Rocheteau, and Weill (2015) provide a detailed analysis of competitive search in the present semi-centralized model. Wright, Kircher, Guerrieri, and Julien (2017) offer a survey of the literature, and Guerrieri and Shimer (2014), Chang (2018), Armenter and Lester (2017), Williams (2014), Chaumont (2018), Li (2018) and Gabovski and Kospentaris (2020) for OTC market applications.

2.3 Dynamic programming

The Hamilton-Jacobi Bellman (HJB) equations for the customer's maximum attainable utilities are:

$$\begin{aligned} rV_0(\delta) &= \quad \gamma \int [V_0(\delta') - V_0(\delta)]\, f(\delta')\, d\delta' + \lambda \max\{\Delta V(\delta) - A(\delta), 0\} \\ rV_1(\delta) &= \delta + \gamma \int [V_1(\delta') - V_1(\delta)]\, f(\delta')\, d\delta' + \lambda \max\{B(\delta) - \Delta V(\delta), 0\}. \end{aligned}$$

Taking the difference between the two equations, and using the expression for the bid and the ask prices in (1) and (2) above, we obtain that $\Delta V(\delta) = V_1(\delta) - V_0(\delta)$ solves:

$$\begin{aligned} r\Delta V(\delta) = \delta + \gamma \int [\Delta V(\delta') - \Delta V(\delta)]\, f(\delta')\, d\delta' + \lambda(1-\theta) \max\{P - \Delta V(\delta), 0\} \\ - \lambda(1-\theta) \max\{\Delta V(\delta) - P, 0\}. \end{aligned}$$

The left-hand side, $r\Delta V(\delta)$, is the annuitized or "flow" reservation value. The right-hand side decomposes this flow reservation value into four components. The first component, δ, is the flow utility received by a customer who holds one unit of asset. The second component, $\gamma \int [\Delta V(\delta') - \Delta V(\delta)]\, f(\delta')\, d\delta$, is the flow of expected net utility of a type change. That is, with intensity γ, the customer draws a new type δ' according to the density $f(\delta')$. This changes the reservation value by $\Delta V(\delta') - \Delta V(\delta)$. The third term, $\lambda(1-\theta)\max\{P - \Delta V(\delta), 0\}$, is the net utility flow of selling assets. The fourth term, $-\lambda(1-\theta)\max\{\Delta V(\delta)-P,0\}$, is minus the net utility flow of purchasing.

A first takeaway from the equation is that reservation values depend on two search options. On the one hand, as shown by the third term, reservation values depend positively on the option to search for a higher selling price, which increases a customer's willingness to pay. On the other hand, as shown by the fourth term, it depends negatively on the option to search for a lower purchasing price, which naturally decreases a customer's willingness to pay. As will become clear, the price impact of search frictions ultimately depends on the relative magnitude of these two option values.

A second takeaway from the HJB equation is that trading through dealers with intensity λ and bargaining power θ is payoff equivalent to trading directly in the centralized inter-dealer market, but with a bargaining-adjusted intensity $\lambda(1-\theta)$. This important observation means that what ultimately matters for valuation is not the true physical search intensity, λ, but the bargaining-adjusted search intensity, $\lambda(1-\theta)$. Hence, even when customers contact dealers relatively quickly, search frictions may have a large impact on valuation if dealers have a sufficiently large bargaining power.

A first explicit solution. Now using that $\max\{P-\Delta V(\delta),0\}-\max\{\Delta V(\delta)-P,0\} = P-\Delta V(\delta)$, the HJB equation can be written:

$$r\Delta V(\delta) = \delta + \gamma \int \left[\Delta V(\delta') - \Delta V(\delta)\right] f(\delta')\, d\delta' + \lambda(1-\theta)\left[P - \Delta V(\delta)\right]. \tag{3}$$

By inspection, one sees that this HJB equation is solved by:

$$\Delta V(\delta) = \mathbb{E}\left[\int_0^\tau e^{-rt}\delta_t\, dt + e^{-r\tau}P \,\middle|\, \delta_0 = \delta\right],$$

where τ is exponentially distributed with bargaining-adjusted intensity $\lambda(1-\theta)$, and δ_t is the stochastic utility flow at time $t>0$, which in general differs from $\delta_0=\delta$, because the customer may have switched type. The formula shows that the customer's reservation value is equal to the expected present value of enjoying utility flows until the next contact time with the inter-dealer market, at which point the value of holding the asset vs. not is simply equal to the inter-dealer price. After a few lines of algebra, one can rewrite this equation as:

$$\Delta V(\delta) = \left(1-\mathbb{E}\left[e^{-r\tau}\right]\right)\frac{D(\delta)}{r} + \mathbb{E}\left[e^{-r\tau}\right]P, \tag{4}$$

where

$$D(\delta) \equiv \frac{\mathbb{E}\left[\int_0^\tau e^{-rt}\delta_t\,dt \,\middle|\, \delta_0 = \delta\right]}{\mathbb{E}\left[\int_0^\tau e^{-rt}\,dt\right]}.$$

Equation (4) shows that the reservation value is a convex combination of two terms. To interpret the first term notice that $D(\delta)$ is the average discounted flow utility of the customer, from $t = 0$ to $t = \tau$. The second term is the price. One sees in particular that, if $\lambda(1-\theta) \to \infty$, then $\mathbb{E}\left[e^{-r\tau}\right] = \frac{\lambda(1-\theta)}{r+\lambda(1-\theta)} \to 1$, and the reservation value converges to the price. This makes sense: if the customer can contact the inter-dealer market instantly, then her reservation value is simply the value of selling the asset on that market at price P, regardless of δ. Hence, the only reason why the reservation value is different from the price is that there are search and bargaining frictions: a customer who owns an asset must hold it for some time before selling it to the market, with an expected discounted valuation given by the first term in the equation.

A second explicit solution. Let us first solve for the average reservation value by taking expectations on both sides of (3) with respect to the probabiliy density $f(\delta)$:

$$r \int \Delta V(\delta') f(\delta')\,d\delta' = \int \delta' f(\delta')\,d\delta' + \lambda(1-\theta)P - \lambda(1-\theta)\int \Delta V(\delta') f(\delta')\,d\delta',$$

which implies that

$$\int \Delta V(\delta') f(\delta')\,d\delta' = \frac{\int \delta' f(\delta')\,d\delta' + \lambda(1-\theta)P}{r + \lambda(1-\theta)}.$$

Substituting back into (3) gives, after a few lines of algebra:

$$\Delta V(\delta) = \frac{r}{r+\lambda(1-\theta)}\left[\frac{r+\lambda(1-\theta)}{r+\gamma+\lambda(1-\theta)}\frac{\delta}{r} + \frac{\gamma}{r+\gamma+\lambda(1-\theta)}\int \frac{\delta'}{r} f(\delta')\,d\delta'\right] + \frac{\lambda(1-\theta)}{r+\lambda(1-\theta)}P.$$

Comparing with (4), one obtains that:

$$D(\delta) = \frac{r+\lambda(1-\theta)}{r+\gamma+\lambda(1-\theta)}\delta + \frac{\gamma}{r+\gamma+\lambda(1-\theta)}\int \delta' f(\delta')\,d\delta'. \tag{5}$$

This is an explicit expression for the average discounted flow utility of the customer from $t = 0$ to $t = \tau$. It shows for example that $D(\delta)$ converges to δ if $r \to \infty$, $\lambda(1-\theta) \to \infty$, or if $\gamma \to 0$. In all these cases, $D(\delta)$ is mostly determined by its initial value, $\delta_0 = \delta$.

2.4 Market clearing: allocations

To write the market-clearing condition, I note that, during a small time interval of length h, the measure of customers in contact with dealers is equal to λh, and recall that these customers are sampled independently from the entire population. This has two key implications.

First, within the small group of customers in contact with dealers, the fraction of asset holders is s, the per capita asset supply in the overall customer's population. Therefore, the gross supply of assets during the small time interval of length h is $\lambda h s$.

Second, within the small group of customers in contact dealers, the distribution of utility flows is $f(\delta)$, the same as in the customers' overall population. From the analysis of reservation values, it then follows that when a customer meets a dealer, her post-trade holding is $q = 1$ if and only if $\Delta V(\delta) > P$. But since the reservation value is strictly increasing in δ, and is unbounded above and below, it implies that there is some cutoff $\delta^\star$ (possibly outside the support) such that $\Delta V(\delta) > P$ if and only if $\delta > \delta^\star$. The customer with utility type $\delta^\star$ is indifferent between holding the asset or not, i.e. $\Delta V(\delta^\star) = P$, and she is commonly referred to as the "marginal customer". Taken together, the gross demand of assets is $\lambda h \int_{\delta \geq \delta^\star} f(\delta')\, d\delta'$.

Equating gross supply and gross demand, I find that $\delta^\star$ solves:

$$s = \int_{\delta \geq \delta^\star} f(\delta')\, d\delta'. \tag{6}$$

This condition reveals that the utility type of the marginal customer, $\delta^\star$, is independent of the search-intensity parameter, λ. However, this does not mean that search frictions do not matter for allocations: it only means that search frictions do not determine who holds the asset, within the small group of customers who are currently in contact with dealers. There are other customers below the marginal type, $\delta < \delta^\star$, who hold the asset because they have switched type and have not been able to contact a dealer. Likewise, there are customers above the marginal type who do not hold.

To see precisely how search frictions matter for allocations, let $\psi_q(\delta)$ denote the density of customers who hold q units of the asset with utility flow δ. In a steady state, the densities must solve the following three equations:

$$\psi_0(\delta) + \psi_1(\delta) = f(\delta) \tag{7}$$

$$\int \psi_1(\delta')\, d\delta' = s \tag{8}$$

$$\lambda\psi_0(\delta)\mathbb{I}_{\{\delta>\delta^\star\}} + \gamma \left[\int \psi_1(\delta')\, d\delta'\right] f(\delta) = \lambda\psi_1(\delta)\mathbb{I}_{\{\delta\leq\delta^\star\}} + \gamma\psi_1(\delta). \tag{9}$$

The first equation, (7), states that the total density of type-δ customers must be equal to $f(\delta)$. The second equation, (8), is a feasibility condition, stating that all assets must be held. The third equation, (9), equates inflow (on the left-side) and outflow (on the right-side) into the set of customers with holding $q = 1$ and utility flow δ. (The counterpart of equation (9) for $q = 0$ is redundant). Combining the three equations gives:

$$\frac{\psi_1(\delta)}{f(\delta)} = \begin{cases} \frac{\gamma}{\lambda+\gamma}s & \text{if } \delta < \delta^\star \\ \frac{\gamma}{\lambda+\gamma}s + \frac{\lambda}{\lambda+\gamma} & \text{if } \delta \geq \delta^\star. \end{cases}$$

The formula confirms that assets are held by customers of all utility types. However, thanks to re-allocation through the search market, customers with utility flow $\delta > \delta^\star$ are more likely to hold assets. In fact, when $\lambda \to \infty$ and the market becomes frictionless, assets are only held by customers with utility flow $\delta > \delta^\star$. When $\lambda \to 0$, assets are held randomly across the utility flow spectrum. For intermediate values of λ, the distribution is a convex combination of the $\lambda \to \infty$ frictionless allocation, and of the $\lambda = 0$ random allocation, with a convex weight $\lambda/(\lambda + \gamma)$. The magnitude of the weight is fully determined by the ratio λ/γ. For instance, if $\gamma \simeq 0$, the allocation converges to the frictionless allocation, even though $\lambda < \infty$. This is because, in this case, customers never change utility flow. As a result, after her first contact time with dealers, a customer with $\delta > \delta^\star$ will hold an asset forever.

2.5 Market clearing: prices

I turn to an analysis of prices. From the previous section, it follows that, conditional on contacting dealers, a customer finds it optimal to hold the asset if and only $\delta > \delta^\star$. Therefore, $P = \Delta V(\delta^\star)$. Plugging this equality into equation (4) evaluated at $\delta = \delta^\star$, it follows that:

$$P = \frac{D(\delta^\star)}{r} = \frac{1}{r}\left(\frac{r + \lambda(1-\theta)}{r + \gamma + \lambda(1-\theta)}\delta^\star + \frac{\gamma}{r + \gamma + \lambda(1-\theta)}\int \delta' f(\delta')\, d\delta'\right), \tag{10}$$

where the second equality follows from (5).

To characterize the impact of search frictions on the asset price, it is useful to analyze two limit cases: when frictions are very large, $\lambda(1-\theta) \to 0$, and when frictions are very small, $\lambda(1-\theta) \to \infty$:

$$\lim_{\lambda(1-\theta)\to 0} P = \mathbb{E}\left[\int_0^\infty e^{-rt}\delta_t\,dt \,\Big|\, \delta_0 = \delta^\star\right] = \frac{1}{r}\left(\frac{r}{r+\gamma}\delta^\star + \frac{\gamma}{r+\gamma}\int \delta'\, f(\delta')\, d\delta'\right). \tag{11}$$

$$\lim_{\lambda(1-\theta)\to\infty} P = \frac{\delta^\star}{r}. \tag{12}$$

Equation (11) is the price when customers expect that it will take them an infinitely long bargaining-adjusted time to contact dealers. The expression reveals that the price in this case is equal to the autarky, or buy-and-hold, utility of the marginal customer. One can easily verify that it also coincides with the asset price if a competitive market opens at time $t = 0$ only and never after.

Equation (12) is the price when customers can trade instantly. In that case, the price is equal to the present value of the utility flow of a hypothetical customer who remains the marginal type forever. This is so even though, according to our specification of the utility flow process, no customer actually remains marginal forever. Indeed, whenever a customer's utility type jumps below $\delta^\star$, he finds it optimal to immediately sell and vice versa when an customer's utility flow jumps above $\delta^\star$. As a result, the market price capitalizes the flow utility of a hypothetical customer who is always indifferent between buying and selling.

Comparing (11) and (12), one sees that the $\lambda(1-\theta) \to 0$ price is smaller than the $\lambda(1-\theta) \to \infty$ if $\delta^\star > \int \delta' f(\delta')\, d\delta'$ and greater otherwise. Correspondingly, when $\lambda(1-\theta)$ increases from zero to infinity, the price increases if $\delta^\star > \int \delta' f(\delta')\, d\delta'$, and it decreases otherwise.

To understand this finding, recall that reservation values depend positively on the search option to sell, and negatively on the option to buy. When $\lambda(1-\theta)$ increases, both options become more valuable, impacting reservation values in opposite ways. But when $\delta^\star > \int \delta' f(\delta')\, d\delta'$ the marginal customer expects her utility flow to fall over time, so the option to sell if she owns has a larger value than the option to buy if she does not. Hence, as $\lambda(1-\theta)$ increases, the positive effect of the option to sell dominates the negative effect of the option to buy, and the price increases.

Feldhütter (2012) argues that this observation helps to empirically identify times of strong selling pressures in the market. To understand his argument within the benchmark model, note that the condition $\delta^\star < \int \delta' f(\delta')\, d\delta'$ indicates a strong selling pressure: indeed, (6) reveals that it holds either when supply is large or when demand is low. As argued above, under this strong selling pressure condition, the option value to sell is less valuable than the option to buy, and so an increase in $\lambda(1-\theta)$ decreases prices. Hence, in a cross section of investors, one should expect that the most sophisticated ones, that is, those with a high $\lambda(1-\theta)$, trade at lower prices than less sophisticated ones. The opposite is true when the selling pressure subsides. Hence, the sign of the price differential between sophisticated and less sophisticated investors helps identify times of strong selling pressure in the market.

Decomposing the yield spread. Imagine that the asset is a fixed-income security, such as a municipal or corporate bond. Then, the natural empirical counterpart of the discount is the liquidity yield spread. That is, let us define the yield on the asset to be the $r+\ell$, where ℓ is the spread over the risk-free rate that makes the present value of the marginal customer's utility flow equal to the price:

$$P = \frac{\delta^\star}{r+\ell}.$$

Notice the desirable property that $\ell = 0$ when there is no friction. Then using equation (10) I obtain that:

$$\frac{\ell}{r+\ell} = \frac{\gamma}{r+\gamma+\lambda(1-\theta)}\left(1 - \frac{\int \delta' f(\delta')\, d\delta'}{\delta^\star}\right).$$

This equation confirms that the liquidity yield spread is positive if and only if $\delta^\star > \int \delta' f(\delta')\, d\delta'$, that is, if and only if a type change reduces the the marginal customer's utility flow for the asset, on average, which may be interpreted as financial distress. The equation also shows that the yield spread depends jointly on several parameters: it is increasing in the frequency of financial distress, measured by γ, in the average distress cost, measured by $1 - \int \delta' f(\delta')\, d\delta'/\delta^\star$, in the size of search frictions, as measured by $1/\lambda$, and in the market power of dealers, as measured by θ. Gavazza (2016) and Hugonnier, Lester, and Weill (2019) discuss how bargaining power and distress cost are separately identified by the level of prices (yield spread) and transaction costs (bid ask spread).

3 Thematic tours of the literature

This section provides thematic tours of the literature. Each tour focuses on a collection of papers sharing a methodological or substantive theme. The tours often cross paths: papers may appear multiple times if they fit in different themes.

3.1 Entry

A number of authors have considered entry decisions in OTC markets. Consider for example the entry decision of dealers. Assume that, if the total measure of dealers in the market is μ_d, then customers' search intensity is given by an increasing and concave function $\lambda(\mu_d)$ satisfying Inada conditions. Hence, when the measure of dealers increase, the semi-centralized market creates more contacts with customers, but becomes more congested. Flow profit per dealer can be shown to be:

$$\frac{\lambda(\mu_d)\gamma s(1-s)}{\mu_d\left(\gamma+\lambda(\mu_d)\right)}\frac{\theta\left(\mathbb{E}\left[\delta'-\delta^\star \,|\, \delta'>\delta^\star\right]+\mathbb{E}\left[\delta^\star-\delta' \,|\, \delta'<\delta^\star\right]\right)}{r+\gamma+\lambda(\mu_d)(1-\theta)},$$

where all expectations are taken with respect to the distribution $f(\delta')$. The first term is the volume per dealer and the second term is the average bid ask spread. One sees that the profit per dealer is decreasing in μ_d, for two reasons. First, although the total volume increases with μ_d, the volume per dealer goes down because $\lambda(\mu_d)$ is convave. Second, the entry of dealers improves the outside option of customers and reduces the average bid ask-spread. Clearly, there is a unique μ_d such as flow profits and entry costs are equalized. One can show that, relative to a utilitarian planner's benchmark, dealers enter too much if θ is too large, and too little if θ too small, as in many other search and matching models.

In his structural study Gavazza (2016) found excessive entry of dealers in the market for commercial aircraft. Vayanos and Wang (2007), Afonso (2010), Rocheteau and Weill (2011), Sambalaibat (2015) considered the entry of customers in OTC markets. Atkeson, Eisfeldt, and Weill (2015) considered both entry and exit of investors with heterogenous valuation, and showed that, depending on their preferences, some investors endogenously choose to enter to and assume the role of dealers, while others choose to enter and assume the role of customers. They show that, under natural conditions, too many investors enter to assume the role of dealers, and too little to assume the role of customers. In Farboodi, Jarosch, and Shimer (2018a) ex ante identical investors endogenously choose different search intensities upon entry, resulting in an equilibrium market structure similar to the one exogenously assumed in the benchmark model. Finally, models of competitive search typically impose a free entry condition as well, but typically obtain efficiency under symmetric information.

3.2 Unrestricted asset holdings

In the benchmark model, we assumed for simplicity that investors can hold either one or zero unit of this asset. Gârleanu (2009) and Lagos and Rocheteau (2009) showed independently that the model can be generalized to the case of unrestricted asset holdings. In particular, Lagos and Rocheteau (2009) consider the following specification of preferences: they assume that an investor with utility type δ who holds q units of the asset derives a flow utility equal to $\delta u(q)$, and the utility function satisfies Inada conditions.[4] More recent work with unrestricted holdings but fully decentralized search markets include Afonso and Lagos (2015a), Cujean and Praz (2013), Üslü (2019), and Üslü and Velioğlu (2019).

With unrestricted asset holdings, the pricing formula of the benchmark model remains valid at the margin, with a utility flow equal to $\delta u'(q)$ instead of just δ. The difference, of course, is that the asset holding q is now endogenous and depends on the magnitude of the frictions. For example, customers endogenously respond to a reduction in frictions by increasing the size of their trades – in this sense, they demand more transaction services from dealers. This can have an important impact on the model's prediction. Consider for example the question of dealers' entry in Lagos and Rocheteau (2007). In the benchmark model, we have seen that per-dealer profits are decreasing in the measure of dealers, so that the entry equilibrium is uniquely determined. In the model with unrestricted asset holdings, there is an additional effect: with more dealers, customers meet dealers more quickly and so are willing to trade larger quantities. This effect can make per-dealer's profit increase in the measure of dealers, creating strategic complementarities in entry decisions and multiple equilibria.

3.3 Alternative assumptions about the matching technology

The matching function is defined as the mapping between the measures of various market participants and the total flow of meetings between them. For example, in the semi-centralized

[4]Without the Inada conditions, the benchmark model with $q \in \{0, 1\}$ asset holding is a special case of the model with unrestricted holdings: indeed, equilibria in which holdings are restricted to $q \in \{0, 1\}$ remain equilibria with unrestricted holdings, as long as utility is Leontief $u(q) = \min\{q, 1\}$ and $q \geq 0$. This is because, in this case, investors find it optimal to hold either $q = 0$ or $q = 1$. With Inada conditions, the benchmark model is no longer a special case but it can be approximated with arbitrary accuracy by appropriate choices of $u(q)$ (see Biais, Hombert, and Weill, 2014).

market, if the measure of customers is $\mu_c = 1$ and the measure of dealers is μ_d, then the flow of meetings is simply $\lambda\mu_c$. More generally, the matching function could depend positively on both the measure of customers and dealers in arbitrary ways. It is customary in the search literature to assume that the matching function has constant returns to scale: that is, scaling up all measures by the same constant scales the flow of meetings linearly, but does not impact search times. Put differently, liquidity does not depend on market size. In labor market applications, this assumption appears to be supported by empirical evidence (Petrongolo and Pissarides, 2001). But in finance, it is not unreasonable to believe that larger markets are more liquid – for example, empirically, there typically exists a positive relationship between issue size and liquidity measures. Correspondingly, some of the literature has considered increasing returns in matching (Pagano, 1989; Vayanos and Wang, 2007; Weill, 2008; Vayanos and Weill, 2008; Shen, Wei, and Yan, 2015; Sambalaibat, 2015).

An and Zheng (2018) and An (2019) have used a different model of the matching process, in the tradition of the stock flow-matching literature (Coles and Smith, 1992). According to this model, trading delays do not arise from search frictions, but instead because customers only demand specific assets. Assume for example that there is a flow γ_b of customers who turn into buyers and instantly contact the market. Each buyer is only willing to purchase a finite collection of assets, drawn at random in the continuum according to a Poisson distribution with parameter ρ, independently across buyers. As a result, the demand for any particular asset will take time to materialize – for example, if there is just one single asset for sale, then a customer willing to buy this asset will arrive at an exponentially distributed time with parameter $\rho\gamma_b$. But this means that, in equilibrium, a stock of buyers and sellers is waiting in the market for suitable counterparties to arrive. In this context, An and Zheng and An show that imperfectly competitive dealers have incentives to hold inventories in spite of high holding costs. In their equilibrium, dealers source assets in two ways: from their inventory, or from the set of sellers waiting for a suitable counterparty. This is appealing because it resembles risky principal and riskless-principal trade in practice.

3.4 Fully decentralized OTC markets

The semi-centralized market assumption is based on the view that inter-dealer frictions are smaller than customer-to-dealer frictions. However, much of the micro data about OTC market concerns inter-dealer trades. For example, in the corporate bond data from the Trade Reporting and Compliance Engine (TRACE), anonymized identifiers track the trades of particular dealers, but no such identifiers are provided for customers. Hence, for better or worse, the inter-dealer market is our main empirical laboratory for studying OTC market frictions. This requires to go beyond semi-centralized markets and formulate models that are fully decentralized, in the sense that all trades, including inter-dealers, occur in a decentralized market. Moreover, the empirical evidence about inter-dealer trades suggests that there is substantial heterogeneity between dealers, in terms of their trading volume, their markups, whether they tend to trade more often with customers or with other dealers, etc...

But solving models with heterogenous dealers in fully decentralized markets turns out to be quite complex. To understand why, assume as in the benchmark model that agents are heterogeneous in terms of their utility flow, but that they must search for counterparties in a fully decentralized market. Then, the reservation value $\Delta V(\delta)$ depends on their expected terms of trade, which in turn depends on the distribution of reservation values across potential counterparties, determined by $\psi_1(\delta)$ and $\psi_0(\delta)$. Hence, one must wrestle with a potentially high dimensional fixed-point problem: characterizing the two-way feedback between this distribution and agents' optimal trading decisions.[5] An active recent literature has developed methods to solve this problem: Afonso and Lagos (2011), Neklyudov (2019), Atkeson, Eisfeldt, and Weill (2015), Colliard and Demange (2014), Hugonnier, Lester, and Weill (2014, 2019), Farboodi, Jarosch, and Shimer (2018a), Üslü (2019), Shen, Wei, and Yan (2015), Bethune, Sultanum, and Trachter (2018), Chang and Zhang (2018), Farboodi, Jarosch, and Shimer (2018a), Farboodi, Jarosch, Menzio, and Wiriadinata (2018b), Tse and Xu (2018), Yang and Zeng (2019), and Üslü and Velioğlu (2019). Because of their rich implications and because they apply to inter-dealer trades, these models take the literature one step closer to structural estimation and ex-ante policy evaluation. In particular, Liu (2020) recently offered a structural estimation of Hugonnier, Lester, and Weill (2019) with endogenouss search intensity. But more progress is needed in this important area of research.

A key theoretical insight in this literature is that, with heterogenous agents and fully decentralized markets, some agents endogenously emerge as dealers, and other as customers. To understand why, consider our benchmark model in which agents have heterogenous utility flows, δ (a similar argument applies to other types of heterogeneity). In a fully decentralized market, agents have heterogenous reservation values and draw counterparties at random from the entire population. Agents whose reservation value $\Delta V(\delta)$ is closer to the economy wide

[5]Semi-centralized markets are simpler because the distribution of reservation values across counterparties is degenerate: indeed, all customers' counterparties are dealers, and all dealers trade in a centralized market, equalizing their reservation values to the price P.

median are equally likely to meet with agents who have either higher or lower reservation value, and so equally likely to be buyers or sellers. In equilibrium, these median agents buy and sell repeatedly, so that their gross trade exceeds their net trade. In this sense, they endogenously emerge as intermediaries or dealers. Agents with extreme reservation values are more likely to trade in just one direction, buy or sell, and so they endogeously emerge as customers.[6]

While different types of heterogeneity can generate similar patterns of endogenous intermediation, they have different economic implications. Perhaps the simplest example would be intermediation arising because of differences in trading speed (Farboodi, Jarosch, and Shimer, 2018a), versus difference in rent-extraction ability (Farboodi, Jarosch, Menzio, and Wiriadinata, 2018b). The former leads to strictly efficient intermediation while the latter does not. Dugast, Üslü, and Weill (2019) show that heterogeneity in private values generates OTC markets that are too small relative to centralized markets, while heterogeneity in trading technology has the opposite implication.

One important open question for this theoretical literature is empirical and quantitative: how to determine the relative importance of different types of heterogeneity. Üslü (2019) and Dugast, Üslü, and Weill (2019) observe for example that the patterns of net and gross trade volume in the cross section can help distinguish between heterogeneity in flow utility, and heterogenetiy in trading technology. But more work remains to be done, by studying the qualitative and quantitative implications of different types of heteregeneity for a broader set of market outcomes.

3.5 Dynamic market response to shocks

A substantial body of empirical evidence suggests that an important dimension of illiquidity is the temporary price impact of supply shocks, and the extent to which intermediaries step in to mitigate these supply shocks by taking inventories. The temporary price impact of supply shocks has been studied by Duffie, Gârleanu, and Pedersen (2007), Feldhütter (2012), Trejos and Wright (2016) and Akın and Platt (2019). The manner in which dealers endogenously respond to supply shock by accumulating inventories and providing liquidity to customers is the topic of Weill (2004, 2007), Lagos, Rocheteau, and Weill (2011), Weill (2011), Di Maggio (2013).

[6]The argument in this paragraph relies on matching being random. However, Chang and Zhang (2018) show that it is not necessary. They consider a dynamic bilateral market in which agents are ex-ante identical and have unobservable valuation. In contrast with the above cited literature, meetings are not random: agents can choose with whom to match. They show that endogenous intermediation arises too, as an efficient coordination mechanism to dynamically reduce misallocation. Gabovski and Kospentaris (2020) also show that intermediation can arise in the absence of random matching, in the context of a competitive search model.

Technically, studying this question requires relaxing the constraint that dealers cannot hold inventories. While this constraint does not bind in the steady state of the benchmark model, it turns out to matter out of steady-state. Dealers have incentives to accumulate inventories when the supply shock hit, because they anticipate that they will be able to resell them more quickly than customers when the shock subsides. Models in this vein shed light on the recent policy debate regarding the potentially negative impact of post-crisis regulation on bond market liquidity (Bao, O'Hara, and Zhou, 2018; Bessembinder, Jacobsen, Maxwell, and Venkataraman, 2018; Dick-Nielsen and Rossi, 2019). Empirical work has highlighted that the answer depends on the liquidity measure considered. For example, some authors have argued that the bid-ask spread has not increased. However, in response to supply shocks such as de-listing or downgrading, dealers accumulate less inventories than before, and they earn higher returns from liquidity provision. This prediction is entirely consistent with search-based models of dealer's liquidity provision, for example a simple extension of Weill (2007) in which dealers have positive bargaining power. According to the model, the reason the bid-ask spread does not go up is that inventory costs move the bid and the ask in the same direction: dealers are less willing to buy, which pushes the bid down, but by the same token they are more willing to sell, which pushes the ask down. In this model inventory costs do not matter for bid-ask spreads, but they do matter for dealers' return and for the amount of liquidity they supply.

Another context in which this question has been of interest is that of "market freeze." This issue became particularly salient during the Great Financial Crisis of 2008 when trading volume collapsed precisely in markets that were plagued by a sudden increase in asymmetric information: namely, in markets for asset backed securities in which investors had serious concerns about collateral quality. Camargo and Lester (2014) and Chiu and Koeppl (2016) have studied theoretically how the market recovers from such asymmetric information shock, and have characterized welfare improving policy interventions, while Zou (2019) considers dynamic information acquisition decisions.

3.6 Asymmetric information

In the benchmark model I assumed that prices are set under symmetric information about flow valuation and asset fundamental value. This is clearly a strong assumption. In fact, OTC markets are typically considered opaque, with pervasive asymmetric information problems: about customers and dealers' private values, about outside trading opportunities, about assets fundamental value and about aggregate order flow. Many policy interventions and regulations have aimed to alleviate asymmetric information problems. To understand the effect of opaqueness and the potential policy responses, many authors have introduced asymmetric information into search models of OTC markets.

Consider first the case of asymmetric information about private flow valuation. Zhu (2011) considers a sequential search problem of a customer contacting a finite number of dealers, with asymmetric information about valuation but also about the history of past quotes from other dealers. In this context, Zhu shows that repeat contact with the same dealer signals that the customer did not receive good quotes from others, and so prompts the dealer to offer worse terms. Zhang (2017) considers long-term relationships when dealers do not observe customers' flow valuation. He shows, among other results, that dealers use delays to screen customers, creating an endogenous distortion above and beyond the delays that may be created by search. Cujean and Praz (2013) consider a trading mechanism that can accommodate private information about private value in a fully decentralized market and study the impact of increased transparency.

The literature has considered asymmetric information about common value as well. One typical assumption, in line with the lemon market literature, is that all assets in the market are finely differentiated. That is, the value of each asset is drawn independently according to some common distribution, and is known by the seller but not by the buyer. This is relevant to the recent financial crisis. One may argue that in the mortgage based security market, securities had finely differentiated collateral pools, and that investors were asymmetrically informed about the quality of these pools. Part of the literature seeks to explain the manner in which such asymmetric information reduces liquidity, leads to market freezes or fire sales, and creates room for welfare improving policy interventions. Work in this area include Guerrieri and Shimer (2014, 2018) and Chang (2018), who use and extend the competitive search framework of Guerrieri, Shimer, and Wright (2010) to asset markets (see also Williams, 2014; Li, 2018). Camargo and Lester (2014) and Chiu and Koeppl (2016) consider dynamics following an adverse selection shocks. Maurin (2018) shows that such setting may display endogenous fluctuation in liquidity. Zou (2019) studies non-stationary equilibria with information acquisition decisions. Lester, Shourideh, Venkateswaran, and Zetlin-Jones (2019) extends the Burdett and Judd (1983) pricing mechanism to the case of asymmetric information. They show that, in order to evaluate the impact of policy or regulatory proposals OTC markets, it is crucial to simultaneously account for imperfect competition and asymmetric information. For example, increasing competition in an OTC market may be either welfare improving or reducing, depending on the degree of asymmetric information.

To address the question of informational efficiency, the market microstructure literature classically considers asymmetric information in a market for a single asset, not many finely differentiated ones. But this creates an important technical difficulty for search models. Indeed, in a fully decentralized market, investors learn about the asset value via their idiosyncratic matching and trading histories. Hence, to characterize equilibrium outcomes, one needs ways to mathematically track investors' increasingly heterogenous learning histories and to characterize the dynamic of the associated distribution of posterior beliefs. Wolinsky (1990) and recently Lauermann, Merzyn, and Gábor (2018) solved this problem by assuming that new traders continously enter the economy, making the distribution of posterior beliefs stationary. Blouin and Serrano (2001) consider non-stationary dynamics in the setup of Wolinsky. Duffie and Manso (2007) and Duffie, Malamud, and Manso (2009) use convolution methods, assuming that agents fully reveal their information in bilateral meetings. Amador and Weill (2006) consider noisy revelation in a Gaussian setting. Matters become more complicated when one imposes the restriction that signals are generated by trades. Duffie, Malamud, and Manso (2014) considers a double auction setting that endogenously results in full revelation. Golosov, Lorenzoni, and Tsyvinski (2014) characterize long-run outcomes in a setting where trade reveals only partial information.

Lester, Shourideh, Venkateswaran, and Zetlin-Jones (2019) consider a semi-centralized market in which customers are perfectly informed but dealers are not. One key insight of their analysis is that a decrease in search frictions can increase asymmetric information and, correspondingly, the bid-ask spread. To gain some intuition consider the following heuristic argument in the benchmark model. Suppose that customers' flow valuation is of the form $v + \varepsilon_i$, where v is a common value component, and ε_i is an idiosyncratic component. Consider a match between a customer who is informed about v and ε_i, and a dealer who is not. Suppose that, by trading, the dealer can learn the reservation value, $\Delta V(\delta)$. Since the dealer already knows the price, P, this signal is observationally equivalent to $D(\delta)$. Substituting $\delta = v + \varepsilon_i$ into (5), we then obtain that the signal acquired by the dealer is of the form $v + \frac{r+\lambda(1-\theta)}{r+\gamma+\lambda(1-\theta)}\varepsilon_i$, reflecting the average valuation of a customer between two contact times with dealers. The common component has a weight of one, because it never changes between two contact times. The idiosyncratic component may change, so it has a smaller weight. But when $\lambda(1-\theta)$ increases and, correspondingly, search frictions decrease, the idiosyncratic component is less likely to change between two contact times, and so it has higher weight. This increases the noise in the signal, reduces the amount of information about v revealed by the trade, leads to more asymmetric information and widens the spread.

Other approaches to common value-asymmetric information include Duffie, Dworczak, and Zhu (2017), who assume that dealers are uncertain about dealers' common market-making cost, and Brancaccio, Li, and Schurhoff (2017) who consider asymmetric information about the aggregate customer's order flow.

3.7 Multiple assets

Much of the empirical evidence about the impact of liquidity on asset prices is cross-sectional: for example, the evidence that risk-adjusted returns are empirically related to liquidity proxies such as turnover, volume, or bid-ask spread (see, for example Amihud, Mendelson, and Pedersen, 2005, Chapter 3), or the evidence that liquidity is related to violation of no-arbitrage relationships (Amihud and Mendelson, 1991). Comparative statics of single-asset models are not appropriate to interpret such evidence: instead, one needs to formulate models in which multiple assets are traded. In particular, one needs to explain why an investor's ability to choose between payoff equivalent assets does not undo an arbitrage relationship or equalize risk-adjusted return differentials.

In Vayanos and Wang (2007) and Weill (2008) investors must choose between markets for indivisible assets. Vayanos and Wang assume that assets are homogenous, but investors differ in their investment horizons, while Weill assumes that investors are homogeneous but assets differ in supply and possibly other characteristics. In equilibrium, investors are indifferent between asset markets. With increasing returns in matching, large supply assets have higher turnover, lower search times, and higher prices. Therefore, in the cross section, there is an increasing relationship between supply and price, the opposite of what would be obtained via comparative statics in a single asset model. Milbradt (2017) also considers asset heterogeneity, but in a semi-centralized market with endogenous customers' search intensity. A key innovation is that asset types are changing, possibly stochastically, over time. A natural example would be the time-to-maturity, the credit rating, or the distance to default of a bond. Milbradt offers new tools and predictions for the joint relationship between volume, prices, and volatility, in the cross section of asset characteristics.

Vayanos and Weill (2008) study the on-the-run phenomenon – the price differential between recent and old Treasury issues with identical time to maturity. They consider a model with two payoff-identical assets and two search markets: one spot market for buying and selling assets and one repo market for borrowing and lending the assets. Investors are not constrained to choose between markets: they receive trading opportunities for either asset. In equilibrium, short-sellers endogenously coordinate to borrow the same asset. Indeed, short-selling activity increases the turnover, reduces search time, and make it easier to locate the asset when unwinding a short position. The model decomposes the yield spread between its various components and explains what makes arbitrage unprofitable. Sambalaibat (2015) shows that the introduction of credit default swap (CDS) can create liquidity spillover in the underlying bond market. This is because the entry of buyers of protection effectively increases supply. With increasing returns in matching, this increases entry on the other side of the market more than one-to-one, so that bonds become easier to sell and have a higher price. These predictions are supported by evidence from the ban on naked CDS trading during the European debt crisis.

In the papers reviewed above, agents are assumed to be risk neutral, and are restricted to hold either zero or one unit of the many assets traded in equilibrium. While analytically convenient, these assumptions are at odds with classical portfolio theory, in which diversification benefits generate a demand for broad portfolios. This poses a clear challenge to search-based models, in which it is typically difficult to study how asset demand is both shaped by diversification and liquidity concerns. In a recent paper, Üslü and Velioğlu (2019) integrate classical portfolio choice and asset pricing theory within a search model of OTC markets. They assume that agents derive a mean-variance utility flow from asset portfolios, and receive random opportunities to trade specific assets. Üslü and Velioğlu take as given asset–specific search times and risk exposures, and derive analytically cross-sectional prices, volume and price impact. They test the model's key predictions in the corporate bond market.

3.8 OTC versus centralized trade

Why do some investors and assets trade in OTC instead of centralized markets? Answering this question matters a great deal for the ex-ante evaluation of policies that promote centralized market trade. Indeed, the unintended consequences of these policies is likely to depend on the underlying drivers of demand for OTC vs. centralized markets.

The literature on this important topic is growing, and often moves away from the search paradigm. This gives more flexibility to introduce a variety of asymmetric information frictions. A branch of the literature provides comparative static analysis, comparing outcomes across exogenously given market structures, some centralized and other decentralized, or across decentralized structures with varying levels of frictions: Biais (1993), Colliard, Foucault, and Hoffmann (2018), Glode and Opp (2019), Malamud and Rostek (2017), Geromichalos and Herrenbrueck (2016), Li and Song (2019), Liu, Vogel, and Zhang (2018) and Vogel (2019). Another branch of the literature formalizes the demand for OTC vs. centralized market by studying the sorting of heterogenous investors across markets. This includes Yavaş (1992), Gehrig (1993), Rust and Hall (2003), Miao (2005), Lee and Wang (2018), and Yoon (2018), Dugast, Üslü, and Weill (2019). A few papers have explored the manner in which decentralized market structures are endogenously offered in equilibrium due to information and price-setting frictions, and may dominate a centralized exchange. See Kawakami (2017), Farboodi, Jarosch, and Shimer (2018a), Babus and Parlatore (2017), and Cespa and Vives (2018). In these works, authors offer economic explanations for the emergence of decentralized trade, with many insights about the impact of policies. But, for the most part, they take some element of the price setting mechanism as given. Thus, it is typically unclear whether decentralized trade emerges because of the price setting mechanism assumption, or because of some primitive informational frictions. What the literature is perhaps missing is a systematic mechanism and information design approach to the emergence of decentralized trading structures.

3.9 Multiplicity and fragility

Search-and-matching models are a natural framework to explore multiple equilibria, in which "liquidity begets liquidity." As is well known since Diamond (1982), this typically can be achieved via increasing returns in matching (Vayanos and Wang, 2007; Vayanos and Weill, 2008; Sambalaibat, 2018), or other mechanisms (Lagos and Rocheteau, 2007; Chiu and Koeppl, 2016; Sultanum, 2018; Nosal, Wong, and Wright, 2019; Yang and Zeng, 2019). Multiple equilibria are appealing for at least two reasons. First, they can help explain phenomena that are difficult to relate to some underlying fundamental characteristics, such as why do payoff equivalent assets have different prices, or why do observationally similar intermediaries play different roles. Second, they provide natural frameworks to address fragility and to uncover feedback loops leading to deterioration of market liquidity.

3.10 Repeat trade and Relationships

A common criticism of the litterature following Duffie, Gârleanu, and Pedersen (2005) is that counterparties do not trade repeatedly. For example, in the benchmark model, customers never contact the same dealer in the μ_d continuum more than once. This is strongly at odds with empirical evidence (Afonso, Kovner, and Schoar, 2014, and many others). One way to address this criticism is to be explicit about imperfect competition amongst a finite number of dealers – indeed, even with search and random matching, atomic dealers will be contacted repeatedly. Repeat contacts are not just mechanical: they do impact terms of trade (Zhu, 2011; An, 2019). But repeat contacts are different from relationships which create economic value. While the study of long-term relationships is relatively undeveloped in OTC market contexts (with some notable exception such as Hendershott, Li, Livdan, and Schürhoff, Forthcoming; Zhang, 2017; Sambalaibat, 2018), they have been explored in the broader search literature. In particular, the canonical labor market model of Mortensen and Pissarides (1994) addresses the formation of relationship between firms and workers. One perhaps important conceptual distinction is that relationships tend to be exclusive in labor markets, but are non-exclusive in OTC markets: customers may establish simultaneous relationships with several dealers (Afonso, Kovner, and Schoar, 2014).

3.11 Search models of centralized exchange

In the last decades, both individual investors and exchanges have made considerable investments to increase trading speed. Since search theory is built on the premise that speed is scarce in the sense that trade is not instantanenous, it is a natural framework to study the demand and supply for speed. Pagnotta and Philippon (2018) consider competition between exchanges who offer investors to trade in semi-centralized search markets, formally similar to the benchmark model of Section 2. Because customers are heterogenous with respect to the distribution from which they draw preference shocks, they have different marginal valuations for speed. To see this, interpret the benchmark model with $\theta = 0$ as an exchange offering a level of speed λ. In that market the ex-ante flow welfare of customers is:

$$W(\lambda, f) = \int \delta \psi_1(\delta)\, d\delta = s\mathbb{E}\left[\delta'\right] + \frac{\lambda s}{\lambda + \gamma}\left(\mathbb{E}\left[\delta' \mid \delta' \geq \delta^\star\right] - \mathbb{E}\left[\delta'\right]\right),$$

where expectations are with respect to the distribution of utility flows, $f(\delta')$. Under the assumptions that $s = 1/2$ and that $f(\delta')$ is symmetric, one easily sees that $\partial W/\partial \lambda$ increases if the distribution $f(\delta')$ becomes riskier in the sense of second-order stochastic dominance. This is because of the second term in the expression of $W(\lambda, f)$: with a higher speed, customers can acquire assets more quickly upon drawing a flow valuation above $\delta^\star$. Therefore, there is complementarity between the trading speed of an exchange and the riskiness of an investor's flow utility distribution, $f(\delta')$. This implies that Betrand competition between exchanges generates horizontal differentiation in trading speed: exchanges offer higher speed to investors with riskier $f(\delta')$. With fixed costs of establishing exchanges, such differentiation is welfare reducing.

Limit-order markets can also be studied using tools from search theory. Rosu (2009) considers the limit order pricing strategy of patient investors who wait for the Poisson arrival of impatient traders on the other side of the market. Biais and Weill (2009) and Biais, Hombert, and Weill (2014) start from the view that limit-orders represent the trading interest of "absent traders, while they attend to business elsewhere" (Harris, 2003, Section 4.4.5). They consider a version of the semi-centralized model in which the parameter λ now represents the intensity with which a given customer monitors the market. The key difference with the semi-centralized model is that customers are now allowed to trade in two ways: immediately with a market order, or potentially with a delay by leaving a limit order. Dugast (2018) studies limit-order market dynamics following the arrival of unscheduled news.

3.12 Broader implications of OTC market frictions

Much of the literature reviewed so far focuses on partial equilibrium analyses, examining one specific OTC market in isolation from the rest of the economy. But an active branch of the literature offers general equilibrium analyses, studying the two-ways feedback between OTC markets and other economic decisions (e.g. a firms' decision to default, the central bank implementation of monetary policy), other markets (e.g., the primary market for bond issuance), or the economy as a whole. This is an important area of research: it shows how the precise modeling of microeconomic frictions in OTC markets can matter for the study of broader economic questions.

The new monetarist tradition, which pre-dates Duffie, Gârleanu, and Pedersen (2005), has considered general equilibrium models with decentralized markets in which goods are exchanged for money and other assets. In the interest of space, I will not review this branch of the literature in depth, but instead refer the reader to Lagos, Rocheteau, and Wright (2017) for a survey, Nosal and Rocheteau (2011) for a book, and Trejos and Wright (2016) for a comparison with the OTC market literature.

Most fixed income markets are OTC, which likely impacts firms' borrowing cost and their capital structure. He and Milbradt (2014) study the interplay between corporate bond liquidity and default by integrating am OTC secondary market into a capital-structure model à-la Leland

(1994). They show in particular that, when lower fundamentals push firms closer to default, the anticipation of post-default OTC market illiquidity depresses price futher, makes it more costly to roll over, and ultimately accelerates default. Chen et al. (2018) and d'Avernas (2017) build on this framework to decompose yield spreads. Other papers sharing a similar corporate financing focus but different channels include Arseneau, Rappoport, and Vardoulakis (2017), Cui and Radde (Forthcoming), Hugonnier, Malamud, and Morellec (2015), Kozlowski (2018), and Bethune, Sultanum, and Trachter (2019), Roh (2019).

Monetary policy is implemented in the over-the-counter-market for Federal Funds. This has been studied in partial equilibrium by Afonso and Lagos (2015a), Afonso and Lagos (2015b), Armenter and Lester (2017), Afonso, Armenter, and Lester (2019), Bech and Monnet (2016), Lagos and Navarro (2019), Wong and Zhang (2019), and in general equilibrium by Bianchi and Bigio (2014), and Bigio and Sannikov (2019). In particular, in Bianchi and Bigio (2014), banks borrow extra reserves in an OTC interbank market to settle payment shocks. Ex ante, banks accumulate a buffer of reserve to reduce their reliance on OTC market, which impacts lending and aggregate economic activity. Malamud and Schrimpf (2017) develop a market setting in which some asset markets are organized OTC with imperfectly competitive intermediaries, resulting in an imperfect monetary-policy pass-through. Another approach to monetary policy is to explicitly consider that money is a means of payment for assets. While Lucas (1990) assumed a competitive asset market, a recent literature has considered OTC asset markets, with new insights into the impact of monetary policy. Works in this area include Geromichalos and Herrenbrueck (2016), Lagos and Zhang (Forthcoming), Lagos and Zhang (2019a), Lagos and Zhang (2019b), Geromichalos, Herrenbrueck, and Salyer (2016), Mattesini and Nosal (2016), and Lebeau (2019).

An emerging literature considers international economics applications. Geromichalos and Jung (2018), Malamud and Schrimpf (2018) and Bianchi, Bigio, and Engel (2018) study the foreign exchange market, while Passadore and Xu (2018) and Chaumont (2018) study the interplay between liquidity and default for sovereign debt, by integrating a secondary OTC market in the model of Eaton and Gersovitz (1981).

3.13 Calibration and structural estimation

Ever since Duffie, Gârleanu, and Pedersen (2007), many authors have combined their theoretical analyses with calibrations. This includes Vayanos and Weill (2008) for the spread between on and off the run bonds, He and Milbradt (2014) Chen, Cui, He, and Milbradt (2018), and d'Avernas (2017) for decompositions of credit spread between default and liquidity components, Passadore and Xu (2018) and Chaumont (2018) for sovereign spreads, Afonso and Lagos (2015b), Armenter and Lester (2017) and Afonso, Armenter, and Lester (2019) and Bianchi and Bigio (2014) for quantitative policy experiments in the Federal Funds market, Pagnotta and Philippon (2018) for a welfare analysis of speed competitions between exchanges, Hugonnier, Lester, and Weill (2019) for a decomposition of the gains from trade in the muni market, Kozlowski (2018) for an quantitative analysis of maturity choice when debt is traded in OTC market.

Comparatively, the structural estimation of search models of OTC markets remain underdeveloped. Perhaps this is because it must overcome at least two significant obstacles: search models must be rich enough to confront the data, and data must be sufficiently rich to confront implications that are unique to search models. Amongst structural studies of OTC market, Feldhütter (2012) identifies times of strong selling pressure in the corporate bond market, Gavazza (2016) studies the welfare impact of intermediaries in the secondary market for commercial aircraft, Brancaccio, Li, and Schurhoff (2017) studies the contribution of experimentation to learn about order flow in the muni bond market, Hendershott, Li, Livdan, and Schürhoff (Forthcoming) study the formation of dealer-client relationships in the corporate bond market and the welfare impact of unbundling trade and non-trade services provided by intermediaries, and Liu (2020) studies the dynamic search process of dealers. Structural estimation of network models include Gofman (2014), Gofman (2017) and Eisfeldt, Herskovic, Rajan, and Siriwardane (2018)

4 Conclusion

Many asset markets have a decentralized OTC structure, for example the markets for fixed-income securities, some derivatives, as well as repos and federal funds. Over the last two decades OTC markets have been analyzed in a large theoretical and empirical literature. In this review article, I have focused my attention on the branch of the literature that studies OTC market through the lens of search-and-matching theory. I have developed a benchmark model to illustrate the key assumptions and economic forces at play in existing work, and I have discussed some of the key insights generated by the literature. Much work remains to be done, in particular to develop structural models that are sufficiently rich to be estimated based on detailed transaction-based microeconomic data.

A Data underlying the supply breakdown in Figure 1

Equity. I obtain the total market capitalization of U.S. equities from the CRSP Stock File Indexes, and the average daily trading volume by exchange from the webbsite of the Securities Industry and Financial Market Association (SIFMA). I classify the Off Exchange volume reported by SIFMA as OTC, and take the rest to be centralized. I then apportion the market capitalization between OTC and centralized according to the corresponding volume share. That is, I assume that the probability of trading in a OTC vs. centralized market is the same for all equity shares.

Treasuries. I obtain data about Treasury outstanding and daily trading volume from SIFMA. I take the view that Treasury trade on electronic platforms is centralized, and otherwise it is OTC. Unlike with equity, I do not use volume share to appportion the Treasury supply outstanding between OTC and centralized exchange. This because existing evidence suggests that different types of Treasuries trade differently: namely, recently issued or "on-the-run" treasuries trade much more, and are more likely to trade electronically (Barclay, Hendershott, and Kotz, 2006). Formally, under the assumption that about 5 percent of Treasury outstanding are on-the-run (Dupont and Sack, 1999), the share of the overall Treasury supply traded in centralized venue can be written:

$$\alpha_{\text{elec}} = 0.05\alpha_{\text{on,elec}} + 0.95\alpha_{\text{off,elec}},$$

where $\alpha_{\text{on,elec}}$ and $\alpha_{\text{off,elec}}$ denote the volume share of electronic trade for both and on-the-run and off-the-run securities.

There is only limited data available to estimate of $\alpha_{\text{on,elec}}$ and $\alpha_{\text{off,elec}}$. The market research firm Greenwhich Associate reports that the share of electronic trade in the Treasury market is about 60 percent. To apportion this volume between on- and off-the-run treasuries, I rely on descriptive statistics found in a Liberty Street blog, on the website of the Federal Reserve Bank of New York. As shown in Table 1, the blog reports volume by trade and security type using recent data from the Trade and Compliance and Reporting Engine (TRACE), which started recording Treasury transactions in 2017. The data breaks down the volume in the inter-dealer-broker (IDB) segment between electronic and manual trade, but it does not do so for the Dealer-to-Dealer (DTC) and the Dealer-to-Customer (DTD) segments.

Segment	Venue	On-the-run	Off-the-run
Inter-Dealer Broker (IDB)	Electronic	190	0
Inter-Dealer Broker (IDB)	Manual	39	36
Dealer-to-Dealer (DTD)	Electronic and Manual	25	10
Dealer-to-Client (DTC)	Electronic and Manual	142	111

Table 1: Average Daily trading volume by trade and security type, between August 1st, 2017 and July 31, 2018, in billions of USD.

Let $V_{i,k,p}$ denote the volume of security $i \in \{\text{on,off}\}$, in segment $k \in \{\text{IDB,DTC,DTD}\}$, and venue $p \in \{\text{elec,manual}\}$. Using these notations, the share of on- and off-the-run treasuries trade electronically can be written:

$$\alpha_{\text{on,elec}} = \frac{V_{\text{on,IDB,elec}} + \alpha_{\text{elec,DTD+DTC,on}} \left(V_{\text{on,DTD}} + V_{\text{on,DTC}}\right)}{V_{\text{on}}} \tag{13}$$

$$\alpha_{\text{off,elec}} = \frac{\alpha_{\text{elec,DTD+DTC,off}} \left(V_{\text{off,DTD}} + V_{\text{off,DTC}}\right)}{V_{\text{off}}}, \tag{14}$$

given that there is no trade of off-the-run securities in the IDB Electronic segment. Therefore, the calculation boils down to finding $\alpha_{\text{on,DTD+DTC,elec}}$ and $\alpha_{\text{off,DTD+DTC,elec}}$, the share of electronic trade for on- and off-the-run treasuries in the DTD and DTC segment. To do so, I have one additional data moment, the Greenwhich Associate estimate that 60 percent of the Treasury volume is traded electronically. In my notation, it writes as:

$$0.6 = \frac{V_{\text{on,IDB,elec}} + \alpha_{\text{on,DTD+DTC,elec}} \left(V_{\text{on,DTC}} + V_{\text{on,DTD}}\right) + \alpha_{\text{off,DTD+DTC,elec}} \left(V_{\text{off,DTC}} + V_{\text{off,DTD}}\right)}{V_{\text{off}} + V_{\text{off}}}.$$

Clearly, this single equation is not sufficient to pin down the two shares $\alpha_{\text{on,DTD+DTC,elec}}$ and $\alpha_{\text{off,DTD+DTC,elec}}$. To make some progress on this question, I let $x \equiv \alpha_{\text{off,DTD+DTC,elec}} / \alpha_{\text{on,DTD+DTC,elec}}$ denote the relative share. One would assume that $x \in [0, 1]$, that is, on-the-run securities are more likely to be traded electronically. Fixing any value of x, I can solve for $\alpha_{\text{on,DTD+DTC,elec}}$ and I obtain:

$$\alpha_{\text{on,DTD+DTC,elec}} = \frac{0.6 \left(V_{\text{off}} + V_{\text{off}}\right) - V_{\text{on,IDB,elec}}}{V_{\text{on,DTD}} + V_{\text{on,DTC}} + x \times \left(V_{\text{off,DTD}} + V_{\text{off,DTC}}\right)}.$$

I pick $x = 0.5$, which may be considered a conservative choice since, in the IDB segment, $x = 0$. Plugging $x = 0.5$ in this expression together with the volume statistics of Table 1, and using (13) and (14) above, I obtain that a fraction $\alpha_{\text{elec}} = 0.26$ of the Treasury supply outstanding is traded electronically.

Other securities. The rest of the securities, Mortgage Related (Agency and non Agency MBS), Corporate, Municipal, Federal Agency, and Asset Backed, trade in OTC markets.

B A Burdett and Judd model of RFQ

Let us consider the stylized representation of an RFQ auction outlined in the text. Each customer sends $n \geq 2$ simultaneous quote requests to randomly chosen dealers, but dealers decline the request with probability π independently from each other. We obtain from the Binomial formula that, conditional on receiving a quote, a customer receives $k \in \{1, 2, \ldots, n\}$ quotes with probability:

$$\psi_k = (1 - \pi^n)^{-1} \binom{n}{k} (1 - \pi)^k \pi^{n-k}.$$

RFQ are not anonymous. To capture this in a simple (albeit extreme) way, we assume that dealers observe the utility flow of customers. In an RFQ, a dealer observes the number of *requests* made by the customer. But since other dealers may decline to respond, a dealer is uncertain about the number of *quotes* received by the customer. In particular, a dealer expects that a customer has a total of k quotes in hand if $k - 1$ of the $n - 1$ other dealers respond, which occurs with probability

$$\phi_k = \binom{n-1}{k-1} (1 - \pi)^{k-1} \pi^{n-k} = \frac{k}{n(1 - \pi)} \binom{n}{k} (1 - \pi)^k \pi^{-k} = \frac{1 - \pi^n}{n(1 - \pi)} k \psi_k = \frac{k \psi_k}{\sum_\ell \ell \psi_\ell}$$

where the first equality follows by expanding the binomial coefficient, the second by recognizing the formula for π_k, and the third by noting that the ϕ_k must add up to one.

Dealers respond to the request by quoting a price (bid or ask). We can, without loss, represent this price as:

$$(1 - m)P + m\Delta V(\delta).$$

As before m is the share of the surplus appropriated by dealers. It also represents the markup charged by dealers on the ask side, if the customer's reservation value exceeds the inter-dealer price, and the markdown on the bid side, if the customer's reservation value is below the inter-dealer price. Suppose for the remainder of this section that the customer is on the ask side, $\Delta V(\delta) > P$. As will be clear, this is without loss because the bid side is symmetric. Let $G(m)$ denote the probability distribution over the m's quoted by dealers. Since a customer prefers a lower m, the probability distribution over the best m is

$$F(m) = \sum_{k=1}^{n} \psi_k \left[1 - (1 - G(m))^k\right].$$

Now consider the other side of the market. The expected profit for a dealer offering a quote $m \in [0, 1]$ to a customer is

$$\begin{aligned}\Pi(m) =& m\left(\Delta V(\delta) - P\right) \\ &\times \sum_{k=1}^{n} \phi_k \left(1 - G(m-)\right)^{k-1} \left\{\sum_{j=0}^{k-1} \binom{k-1}{j} \left[\frac{G(m) - G(m-)}{1 - G(m-)}\right]^j \left[\frac{1 - G(m)}{1 - G(m-)}\right]^{k-1-j} \frac{1}{j+1}\right\}\end{aligned}$$

where $G(m-) = \lim_{\tilde{m}\uparrow m} G(\tilde{m})$. The term on the first line of the formula, $m(\Delta V(\delta) - P)$, is the profit conditional on the ask quote being accepted. The term on the second line is the probability that the quote is accepted. First, with probability ϕ_k, the offer is received by a customer who has (in total), k simultaneous offers in hand. The probability that all other simultaneous offers are weakly greater than m is $(1 - G(m-))^k$. Conditional on an offer being greater than m, the offer is exactly equal to m with probability $\frac{G(m)-G(m-)}{1-G(m-)}$, and it is strictly greater than m with probability $\frac{1-G(m)}{1-G(m-)}$. Hence, the probability of j other offers equal to m and $k - 1 - j$ other offers strictly greater than m is given by the Binomial formula

$$\binom{k-1}{j} \left[\frac{G(m) - G(m-)}{1 - G(m-)}\right]^j \left[\frac{1 - G(m)}{1 - G(m-)}\right]^{k-1-j}.$$

If there are j other offers equal to m, the customer picks at random and the dealer sells with probability $1/(j + 1)$.

Equilibrium. An equilibrium of the quote-setting game is a distribution of quotes $G(m)$ such that, $\forall m \in \text{support}(G)$, we have that $m \in \underset{m'}{\operatorname{argmax}}\, \Pi(m')$. We first establish, that:

Lemma 1. *If $\psi_1 < 1$ then, in equilibrium, either $G(m)$ is concentrated at $m = 0$, or $G(m)$ is continuous over the support $[\underline{m}, 1]$, for some $\underline{m} > 0$.*

Indeed, suppose that $G(m)$ is not concentrated at $m = 0$. Then, there must exist some $m > 0$ such that $G(m) > 0$. Hence, dealers make strictly positive profit. Since quotes close to 0 would lead to vanishingly small profits, such quote cannot be in the support of the distribution. Hence, support of the distribution must have a strictly positive lower bound, $\underline{m} > 0$. Second, we argue that $G(\cdot)$ must be continuous, that is, it cannot have any mass point. We proceed by contradiction. Suppose that there is discontinuity at some $m > 0$, i.e. $G(m)-G(m-) > 0$. Then it cannot be optimal to quote m. To see why, notice that by reducing m by some small positive amount, which is possible because $m > 0$, a dealer makes strictly larger profits: it eliminates all ties with other firms posting m and so it trades with discretely higher probability, while only making a small profit loss. Third, we argue that there support of the distribution has no gap – mathematically, the support must be the entire interval $[\underline{m}, 1]$. As before we proceed by contradiction. Consider some $m_2 \in (\underline{m}, 1]$ and let $m_1 = \inf\{m < m_2 : G(m) = G(m_2)\}$. Since $G(m)$ is continuous, we have that $G(m_1) = G(m_2)$. Now suppose there is a gap in the distribution m_1 and m_2: mathematically, $m_1 < m_2$ Then any dealer posting a quote in a left neighborhood of m_1 will find it strictly better to post m_2. Indeed, this creates a small decrease in trade probability but a discretely large increase in profit conditional on trading.

Next, we show that:

Proposition 1. *There is a unique equilibrium $G(m)$.*

1. *Monopolistic dealers. If $\psi_1 = 1$, then $G(m)$ is concentrated at $m = 1$;*
2. *Competitive dealers. If $\psi_1 = 0$, then $G(m)$ is concentrated at $m = 0$;*
3. *In between. If $\psi_1 \in (0, 1)$, $G(m)$ is continuous over the support $[\phi_1, 1]$. Moreover, for each m, $G(m)$ is the unique solution of the equation $\phi_1 = m \sum_{k=1}^{n} \phi_k (1-x)^{k-1}$.*

In all cases, the average transaction price for a customer with utility flow δ is $\psi_1 \Delta V(\delta) + (1 - \psi_1)P$.

In the first case, dealers are monopolistic and quote price equal to the reservation value. In the second case dealers are competitive because customers always receive at least two simultaneous quotes. Then, we show by contradiction that $G(\cdot)$ is concentrated at 0. Suppose it is not. Then, from the previous proposition, we know $G(\cdot)$ is continuous over some support $[\underline{m}, 1]$, A dealer's profit is zero when it offers 1, since the customers always receive a strictly better quote with probability one. At the same time, the dealer's profit is strictly positive when it offers $\underline{m}$, since the customer accept with probability one and $\underline{m} > 0$. Hence, the dealer is not indifferent between $m = \underline{m}$ and $m = 1$, and we have reached a contradiction.

Now consider the third case. A distribution concentrated at 0 cannot be an equilibrium: a dealer would find it strictly better to quote $m = 1$, which would yield strictly positive profit in case it is the only quote received by the customers, rather than $m = 0$, which yield zero profits. Hence, in the third case, we know that the distribution of quotes is continuous over some support $[\underline{m}, 1]$. We can go a step further and characterize the quote distribution. Since dealers' profits are equalized for all $m \in [\underline{m}, 1]$:

$$\Pi(1) = \phi_1 \left(\Delta V(\delta) - P\right) = m \left(\Delta V(\delta) - P\right) \sum_{k=1}^{n} \phi_k \left(1 - G(m)\right)^{k-1}.$$

Canceling $\Delta V(\delta) - P$ from both sides we obtain

$$\phi_1 = m \sum_{k=1}^{n} \phi_k \left[1 - G(m)\right]^{k-1}$$

When $m = \underline{m}$, $G(\underline{m}) = 0$, implying that $\underline{m} = \phi_1$. For $m \in (\phi_1, 1)$, the function $G(m)$ is the unique solution of $\phi_1 = m \sum_{k=1}^{\infty} \phi_k (1-x)^{k-1}$, as stated in the proposition.

Finally, we calculate the average transaction price for customers. To do so it is helpful to appeal (informally) to the law of large numbers and notice that the expected net utility of a customer is equal to the average net utility in a measure-one cross section of customers. In addition, the average net utility is equal to the surplus, minus the average dealer's profit. But the average dealer's profits are

$$\left(\sum \ell \psi_\ell\right) \times \phi_1 \left[\Delta V(\delta) - P\right] = \psi_1 \left[\Delta V(\delta) - P\right].$$

Indeed, dealers collectively respond to a total of $\sum \ell \psi_\ell$ quote requests from customers of type δ, and the expected profit per request is ϕ_1. The last equality then follows from the definition of ϕ_1. Taken together, we obtain the stated result.

The brute force calculation is also instructive and useful. The best m received by the customer is the lowest one offered by dealers, with CDF $\sum_{k=1}^{n} \psi_k \left\{1 - \left[1 - G(m)\right]^k\right\}$. The average of these best m is:

$$\begin{aligned}
\int_{\phi_1}^{1} m \sum_{k=1}^{n} \psi_k \, d\left\{1 - \left[1 - G(m)\right]^k\right\} &= \int_{\phi_1}^{1} \sum_{k=1}^{n} \psi_k k \left[1 - G(m)\right]^{k-1} m \, dG(m) \\
&= \left(\sum_{\ell} \psi_\ell \ell\right) \int_{\phi_1}^{1} \sum_{k=1}^{n} \phi_k \left[1 - G(m)\right]^{k-1} m \, dG(m) \\
&= \left(\sum_{\ell} \psi_\ell \ell\right) \phi_1 = \psi_1,
\end{aligned}$$

where the equality on the second line follows because, by its definition, $\phi_k = \psi_k k / \left(\sum_\ell \psi_\ell \ell\right)$, and the equality on the third line follows from the firm's indifference condition.

References

Gara Afonso. Liquidity and congestion. *Journal of Financial Intermediation*, 20:324–360, 2010.

Gara Afonso and Ricardo Lagos. Trade dynamics in the federal funds market. Working paper, Federal Reverve Bank of New York and NYU, 2011.

Gara Afonso and Ricardo Lagos. Trade dynamics in the market for federal funds. *Econometrica*, 83:263–313, 2015a.

Gara Afonso and Ricardo Lagos. The over-the-counter theory of the fed funds market: A primer. *Journal of Money, Credit and Banking*, 47:127–154, 2015b.

Gara Afonso, Anna Kovner, and Antoinette Schoar. Trading partner in the interbank market. Federal Reserve Bank of New York Staff Report no. 620, 2014.

Gara Afonso, Roc Armenter, and Benjamin Lester. A model of the federal funds market: Yesterday, today, and tomorrow. *Review of Economic Dynamics*, Forthcoming, 2019.

Ş Nuray Akın and Brennan C. Platt. Transition dynamics in equilibrium search. Working paper BYU, 2019.

Manuel Amador and Pierre-Olivier Weill. Learning by matching. Working paper UCLA, 2006.

Yakov Amihud and Haim Mendelson. Liquidity, maturity, and the yield on U.S. treasury securities. *Journal of Finance*, 46:479–486, 1991.

Yakov Amihud, Haim Mendelson, and Lasse H. Pedersen. Liquidity and asset prices. *Foundations and Trends in Finance*, 4:270–364, 2005.

Yu An. Competing with inventory in dealership markets. Working paper, John Hopkins, 2019.

Yu An and Zeyu Zheng. Conflicted immediacy provision. Working paper, John Hopkins, 2018.

Roc Armenter and Benjamin Lester. Excess reserves and monetary policy implementation. *Review of Economic Dynamics*, 23(212–235), 2017.

David M. Arseneau, David E. Rappoport, and Alexandros P. Vardoulakis. Private and public liquidity provision in otc markets. Working paper, Federal Reserve Board, 2017.

Andrew Atkeson, Andrea Eisfeldt, and Pierre-Olivier Weill. The market of otc derivatives. Technical report, Working paper, UCLA, 2013.

Andrew Atkeson, Andrea Eisfeldt, and Pierre-Olivier Weill. Entry and exit in otc derivatives markets. *Econometrica*, 83(6):2231–2292, 2015.

Christoph Aymanns, Georg Co-Pierre, and Benjamin Golub. Illiquidity spirals in coupled over-the-counter markets. 2018.

Ana Babus. Market for financial innovation. Working paper, WUSTL and UVA, 2019.

Ana Babus and Maryam Farboodi. The hidden cost of strategic opacity. Working paper, WUSTL and MIT, 2018.

Ana Babus and Tai-Wei Hu. Endogenous intermediation in over-the-counter markets. *Journal of Financial Economics*, 125:200–215, 2018.

Ana Babus and Peter Kondor. Trading and information diffusion in otc markets. *Econometrica*, 86:1727–1769, 2018.

Ana Babus and Cecilia Parlatore. Strategic fragmented markets. Working paper, 2017.

Ana Babus and Cecilia Parlatore. Strategic fragmented markets. Working paper, UW and NYU Stern, 2018.

Jack Bao, Maureen O'Hara, and Alex Zhou. The volcker rule and market making in times of stress. *Journal of Financial Economics*, 130:95–113, 2018.

Michael J. Barclay, Terrence Hendershott, and Kenneth Kotz. Automation versus intermediation: Evidence from treasuries going off the run, working paper. *Journal of Finance*, 61: 2395–2414, 2006.

Morten Bech and Cyril Monnet. A search-based model of the interbank money market and monetary policy implementation. *Journal of Economic Theory*, 164:32–67, 2016.

Hendrik Bessembinder and William Maxwell. Markets: Transparency and the corporate bond market. *Journal of Economic Perspectives*, 22:217–234, 2008.

Hendrik Bessembinder, Stacey Jacobsen, William Maxwell, and Kumar Venkataraman. Capital commitment and illiquidity in corporate bonds. *Journal of Finance*, 73:1615–1661, 2018.

Zachary Bethune, Bruno Sultanum, and Nicholas Trachter. An information-based theory of financial intermediation. Technical report, UVA and Federal Reserve Bank of Richmond, 2018.

Zachary Bethune, Bruno Sultanum, and Nicholas Trachter. Asset issuance in over-the-counter markets. *Review of Economic Dynamics*, 33:4–29, 2019.

Bruno Biais. Price formation and equilibrium liquidity in fragmented and centralized markets. *Review of Financial Studies*, 48:157–185, 1993.

Bruno Biais and Richard C. Green. The microstructure of the bond market in the 20th century. *Revew of Economic Dynamics*, 2019.

Bruno Biais and Pierre-Olivier Weill. Liquidity shocks and order book dynamics. Working paper, TSE and UCLA, 2009.

Bruno Biais, Johan Hombert, and Pierre-Olivier Weill. Equilibrium pricing and trading volume under preference uncertainty. *Review of Economic Studies*, 81:1401–1437, 2014.

Javier Bianchi and Saki Bigio. Banks, liquidity management, and monetary policy. Working paper, NBER 20490, 2014.

Javier Bianchi, Saki Bigio, and Charles Engel. Payments, liquidity, and exchange rates. Working paper, UCLA, 2018.

Saki Bigio and Yuliy Sannikov. A model of credit, money, interest, and price. Working paper UCLA, 2019.

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

Giulia Brancaccio, Dan Li, and Norman Schurhoff. Learning by trading: The case of the u.s. market for municipal bond. Working paper, Cornell University, 2017.

Kenneth Burdett and Kenneth L. Judd. Equilibrium price dispersion. *Econometrica*, 51(4): 955–969, 1983.

Braz Camargo and Ben Lester. Trading dynamics in decentralized markets with adverse selection. *Journal of Economic Theory*, 153:534–568, 2014.

Giovanni Cespa and Xavier Vives. Exchange competition, entry, and welfare. Working paper, 2018.

Briana Chang. Adverse selection and liquidity distorsion. *Review of Economic Studies*, 85: 275–306, 2018.

Briana Chang and Shengxing Zhang. Endogenous market making and network formation. Working paper UW and LSE, 2018.

Gaston Chaumont. Sovereign debt, default risk, and the liquidity of government bonds. Working paper, University of Rochester, 2018.

Hui Chen, Rui Cui, Zhiguo He, and Konstantin Milbradt. Quantifying liquidity and default risks of corporate bonds over the business cycle. *Review of Financial Studies*, 31:852–897, 2018.

Jonathan Chiu and Thor Koeppl. Trading dynamics with adverse selection and search: Market freeze, intervention and recovery. *Review of Economic Studies*, 83:969–1000, 2016.

Melvyn G. Coles and Eric Smith. Marketplaces and matching. *International Economic Review*, 39:239–254, 1992.

Jean-Edouard Colliard and Gabrielle Demange. Asset dissemination through dealer markets. Working Paper, HEC and PSE, 2014.

Jean-Edouard Colliard, Thierry Foucault, and Peter Hoffmann. Inventory management, dealers' connections, and prices in otc markets. Working paper, 2018.

Wei Cui and Sören Radde. Search-based endogenous asset liquidity and the macroeconomy. *Journal of the European Economic Association*, Forthcoming.

Julien Cujean and Rémy Praz. Asymmetric information and inventory concerns in over-the-counter markets. Working paper, University of Bern, 2013.

Adrien d'Avernas. Disentangling credit spreads and equity volatility. Working paper, Stockholm School of Economics, 2017.

Marco Di Maggio. Market turmoil and destabilizing speculation. Working paper, MIT, 2013.

Peter A. Diamond. A model of price adjustment. *Journal of Economic Theory*, 3(2):156–168, 1971.

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

Jens Dick-Nielsen and Marco Rossi. The cost of immediacy for corporate bonds. *Review of Financial Studies*, 32:1–41, 2019.

Darrel Duffie, Semyon Malamud, and Gustavo Manso. Information percolation with equilibrium search dynamics. *Econometrica*, 2009. Forthcoming.

Darrell Duffie. Presidential address: Asset price dynamics with slow-moving capital. *Journal of Finance*, 65:1237–1267, 2010.

Darrell Duffie and Gustavo Manso. Information diffusion in large population. *American Economic Review, Papers and Proceedings*, 97, 2007.

Darrell Duffie, Nicolae Gârleanu, and Lasse H. Pedersen. Securities lending, shorting, and pricing. *Journal of Financial Economics*, 66:307–339, 2002.

Darrell Duffie, Nicolae Gârleanu, and Lasse H. Pedersen. Over-the-counter markets. *Econometrica*, 73:1815–1847, 2005.

Darrell Duffie, Nicolae Gârleanu, and Lasse H. Pedersen. Valuation in over-the-counter markets. *Review of Financial Studies*, 20:1865–1900, 2007.

Darrell Duffie, Semyon Malamud, and Gustavo Manso. Information percolation in segmented markets. *Journal of Economic Theory*, 153:1 – 32, 2014. doi: https://doi.org/10.1016/j.jet.2014.05.006.

Darrell Duffie, Piotr Dworczak, and Haoxiang Zhu. Benchmarks in search markets. *Journal of Finance*, 72:1983–, 2017.

Jérôme Dugast. Unscheduled news and market dynamics. *Journal of Finance*, 78:2537–2586, 2018.

Jérôme Dugast, Semih Üslü, and Pierre-Olivier Weill. A theory of participation in otc and centralized markets. Working paper, U Paris Dauphine, John Hopkins, and UCLA, 2019.

Dominique Dupont and Brian Sack. The treasury securities market: Overview and recent developments. *Federal Reserve Bulletin*, December:785–806, 1999.

Jonathan Eaton and Mark Gersovitz. Debt with potential repudiation: theoretical and empirical analysis. *Review of Economic Studies*, 48:289–309, 1981.

Andrea Eisfeldt, Bernard Herskovic, Srimam Rajan, and Emil N. Siriwardane. Otc intermediaries. Working paper, UCLA, 2018.

Marham Farboodi, Gregor Jarosch, and Robert Shimer. The emergence of market structure. Working paper, MIT, Princeton, University of Chicago, 2018a.

Maryam Farboodi. Intermediation and voluntary exposure to counterparty risk. Working paper, MIT, 2014.

Maryam Farboodi, Gregor Jarosch, Guido Menzio, and Ursula Wiriadinata. Intermediation as rent extraction. Technical report, MIT, Princeton, NYU, and IMF, 2018b.

Peter Feldhütter. The same bond at different prices: Identifying search frictions and selling pressures. *Review of Financial Studies*, 25(4):1155–1206, 2012.

Jean-David Fermanian, Olivier Guéant, and Jiang Pu. The behavior of dealers and clients on the european corporate bond market: The case of multi-dealer-to-client platforms. *Market Microstructure and Liquidity*, 2, 2016.

Miroslav Gabovski and Ioannis Kospentaris. Intermediation in over-the-counter markets with price transparency. Working paper, University of Hawaii Manoa and Virginia Commonwealth University, 2020.

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

Alessandro Gavazza. An empirical equilibrium model of a decentralized asset market. *Econometrica*, 84:1755–1798, 2016.

Thomas Gehrig. Intermediation in search markts. *Journal of Economics and Management Strategy*, 2:97–120, 1993.

Athanasios Geromichalos and Lucas Herrenbrueck. Monetary policy, asset prices and liquidity in over-the-counter markets. *Journal of Money, Credit, and Banking*, pages 35–79, 2016.

Athanasios Geromichalos and Kuk Mo Jung. An over-the-counter approach to the forex market. *International Economic Review*, 59:859–905, 2018.

Athanasios Geromichalos, Lucas Herrenbrueck, and Kevin Salyer. A search-theoretic model of the term premium. *Theoretical Economics*, 11:897–935, 2016.

Sergei Glebkin, Ji Shen, and Bart Zhou Yueshen. Simultaneous multilateral search. Working paper INSEAD, 2019.

Vincent Glode and Christian C. Opp. Over-the-counter vs. limit-order markets: The role of traders' expertise. *Review of Financial Studies*, Forthcoming, 2019.

Michael Gofman. A network-based analysis of over-the-counter markets. Working paper, University of Rochester, 2014.

Michael Gofman. Efficiency and stabilitgy of a financial architecture with too-interconnected-to-fail institutions. *Journal of Financial Economics*, 124:113–14, 2017.

Mikhail Golosov, Guido Lorenzoni, and Aleh Tsyvinski. Decentralized trading with private information. *Econometrica*, 82:1055–1091, 2014.

Veronica Guerrieri and Robert Shimer. Dynamic adverse selection: A theory of illiquidity, fire sales, and flight to quality. *American Economic Review*, 104:1875–1908, 2014.

Veronica Guerrieri and Robert Shimer. Markets with multidimensional private information. *American Economic Journal: Microeconomics*, 85:1502–1542, 2018.

Veronica Guerrieri, Robert Shimer, and Randall Wright. Adverse selection in competitive search equilibrium. *Econometrica*, 78:1823–1862, 2010.

Larry Harris. *Trading and Exchanges: Market Microstructure for Practitioners*. Oxford University Press, New York, New York, 2003.

Frank Hatheway, Amy Kwan, and Hui Zheng. An empirical analysis of market segmentation in u.s. equity markets. *Journal of Financial and Quantitative Analysis*, 52:2399–2427, 2017.

Zhiguo He and Konstantin Milbradt. Endogenous liquidity and defaultable bonds. *Econometrica*, 82:1443–1508, 2014.

Terrence Hendershott, Dan Li, Dmitry Livdan, and Norman Schürhoff. Relationship Trading in OTC Markets. *Journal of Finance*, Forthcoming.

Julien Hugonnier. Speculative behavior in decentralized markets. *Working Paper, Swiss Finance Institute*, 2012.

Julien Hugonnier, Benjamin Lester, and Pierre-Olivier Weill. Heterogeneity in decentralized asset markets. Working paper, EPFL, Federal Reserve Bank of Philadelphia, and UCLA, 2014.

Julien Hugonnier, Semyon Malamud, and Erwan Morellec. Capital supply uncertainty, cash holdings, and investment. *Review of Financial Studies*, 28:391–445, 2015.

Julien Hugonnier, Ben Lester, and Pierre-Olivier Weill. Frictional intermediation in Over-the-Counter markets. *Review of Economic Studies*, Forthcoming, 2019. Working paper EPFL, Federal Reserve Bank of Philadelphia, and UCLA.

ISDA. Key trends in the size and composition of otc derivatives markets. Technical report, International Swap and Derivatives Association, 2018.

Kei Kawakami. Welfare consequences of information aggregation and optimal market size. *American Economic Journal: Microeconomics*, 9(4):303–323, 2017.

Julian Kozlowski. Long-term finance and investment with frictional asset markets. Working paper, Federal Reserve Bank of Saint Louis, 2018.

Ricardo Lagos and Gastón Navarro. Monetary operating procedures in the fed funds market: Theory and policy analysis. Working paper, NYU, 2019.

Ricardo Lagos and Guillaume Rocheteau. Search in asset markets: Market structure, liquidity, and welfare. *American Economic Review P&P*, 97:198–202, 2007.

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

Ricardo Lagos and Shengxing Zhang. Monetary exchange in bilateral over-the-counter markets. *Review of Economic Dynamics*, 2019a.

Ricardo Lagos and Shengxing Zhang. The limit of onetary economics: on money as a medium of exchange in near-cashless economies. Working paper, NYU and LSE, 2019b.

Ricardo Lagos and Shengxing Zhang. Turnover liquidity and the transmission of monetary policy. *American Economic Review*, Forthcoming.

Ricardo Lagos, Guillaume Rocheteau, and Pierre-Olivier Weill. Crises and liquidity in otc markets. *Journal of Economic Theory*, 146:2169–2205, 2011.

Ricardo Lagos, Guillaume Rocheteau, and Randall Wright. Liquidity: A new monetarist perspective. *Journal of Economic Literature*, 55:371–440, 2017.

Stephan Lauermann, Wolfram Merzyn, and Virág Gábor. Learning and price discovery in a search model. *Review of Economic Studies*, 85:1159–1192, 2018.

Lucie Lebeau. Credit frictions and participation in otc markets. Working paper, UCI, 2019.

Tomy Lee and Chaojun Wang. Why trade over the counter? when investors want price discrimination. Working paper, 2018.

Hayne E. Leland. Corporate debt value, bond covenants, and optimal capital structure. *Journal of Finance*, 49:1213–1243, 1994.

Benjamin Lester, Guillaume Rocheteau, and Pierre-olivier Weill. Competing for order flow in otc markets. *Journal of Money, Credit and Banking*, 47(S2):77–126, 2015.

Benjamin Lester, Ali Shourideh, Venky Venkateswaran, and Ariel Zetlin-Jones. Market-making with search and information frictions. Working paper, Federal Reserve Bank of Philadelphia, 2018.

Benjamin Lester, Ali Shourideh, Venky Venkateswaran, and Ariel Zetlin-Jones. Screening and adverse selection in frictional markets. *Journal of Political Economy*, 127(1):338–377, 2019.

Dan Li and Norman Schürhoff. Dealer networks. *Journal of Finance*, 74:91–144, 2019. Working Paper, HEC Lausanne.

Qi Li. Securitization and liquidity creation in markets with adverse selection. Working paper, Penn State University, 2018.

Wei Li and Zhaogang Song. Dealers as information intermediaries in over-the-counter markets. Working paper, 2019.

Shuo Liu. Dealer's search intensity in u.s. corporate bond markets. Working paper, UCLA, 2020.

Ying Liu, Sebastian Vogel, and Yuan Zhang. Electronic trading in otc markets vs. centralized exchange. Working paper, 2018.

Robert E. Jr Lucas. The effects of monetary shocks when prices are set in advance. Reprinted in Lucas (2013), 1989.

Robert E. Jr. Lucas. Liquidity and interest rates. *Journal of Economic Theory*, 50:237–264, 1990.

Robert E. Jr Lucas. *Collected papers on monetary theory*. Harvard University Press, 2013.

Semyon Malamud and Marzena Rostek. Decentralized exchange. *American Economic Review*, 107:3320–3362, 2017.

Semyon Malamud and Andreas Schrimpf. Intermediation markups and monetary policy pass-through. Working paper, EPFL and BIS, 2017.

Semyon Malamud and Andreas Schrimpf. An intermediation-based model of exchange rates. Working paper, EPFL and BIS, 2018.

Mihai Manea. Intermediation and resale in networks. *Journal of political Economy*, 126:1250–1301, 2018.

Fabrizio Mattesini and Ed Nosal. Liquidity and asset prices in a monetary model with otc asset markets. *Journal of Economic Theory*, 164:187 – 217, 2016. doi: https://doi.org/10.1016/j.jet.2015.11.001.

Vincent Maurin. Liquidity fluctuations in over-the-counter markets. Working paper, Stockholm School of Economics, 2018.

Jianjun Miao. A search model of centralized and decentralized trade. *Review of Economic Dynamics*, forthcoming, 2005.

Konstantin Milbradt. Asset heterogeneity in over-the-counter markets. Working paper, Kellogg School of Management, 2017.

Dale T. Mortensen and Christopher A. Pissarides. Job creation and job destruction in the theory of unemployment. *The Review of Economic Studies*, 61(3):397–415, 1994.

Artem Neklyudov. Bid-ask spreads and the over-the-counter interdealer markets: Core and peripheral dealers. *Review of Economic Dynamics*, 33:57–84, 2019.

Yigal S. Newman and Michael A. Rierson. Illiquidity spillovers: Theory and evidence from european telecom bond issuance. Stanford University, 2003.

Ed Nosal and Guillaume Rocheteau. *Money, Payments and Liquidity.* MIT Press, 2011.

Ed Nosal, Yuet-Yee Wong, and Randall Wright. Intermediation in markets for goods and markets for assets. Working paper, Federal Reserve Bank of Atlanta, 2019.

Marco Pagano. Trading volume and asset liquidity. *Quarterly Journal of Economics*, 104: 255–274, 1989.

Emiliano S. Pagnotta and Thomas Philippon. Competing on speed. *Econometrica*, 86(3): 1067–1115, 2018.

Juan Passadore and Yu Xu. Illiquidity in sovereign debt markets. Working paper, EIEF, 2018.

Barbara Petrongolo and Christopher A. Pissarides. Looking into the black box: A survey of the matching function. *Journal of Economic Literature*, 39(2):390–431, 2001.

Rémy Praz. Equilibrium asset pricing with both liquid and illiquid markets. Working paper EPFL, 2015.

Lynn Riggs, Esen Onur, David Reiffen, and Haoxiang Zhu. Swap trading after dodd-frank: Evidence from index cds. Working paper, CFTC and MIT, 2018.

Guillaume Rocheteau and Pierre-Olivier Weill. Liquidity in frictional asset markets. *Journal of Money, Credit, and Banking*, 43:261–282, 2011.

Hee Su Roh. Repo specialness in the transmission of quantitative easing. Working paper, Stanford University, 2019.

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

Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50:97–109, 1982.

John Rust and George Hall. Middlemen versus market makers: A theory of competitive exchange. *Journal of Political Economy*, 111(2):pp. 353–403, 2003.

Batchimeg Sambalaibat. A theory of liquidity spillover between bond and cds markets. Working paper, 2015.

Batchimeg Sambalaibat. Endogenous specialization in dealer networks. Working paper, Kelley School of Business, 2018.

Ji Shen, Bin Wei, and Hongjun Yan. Financial intermediation chains in an otc market. Working paper, DePaul University, 2015.

SIFMA. U.S. fixed income market structure primer. Technical report, Securities Industry and Financial Markets Association, 2018.

Bruno Sultanum. Financial fragility and otc markets. *Journal of Economic Theory*, 177:618–658, 2018.

Alberto Trejos and Randall Wright. Search-based models of money and finance: An integrated approach. *Journal of Economic Theory*, 164:10 – 31, 2016. doi: https://doi.org/10.1016/j.jet.2014.11.005.

Chung-Yi Tse and Yujing Xu. Inter-dealer trades in otc markets – who buys and who sells? Working paper, University of Hong-Kong, 2018.

Laura Tuttle. Otc trading: Description of non-ats trading in national market system stock. Technical report, U.S. Securities and Exchange Commission, 2014.

Semih Üslü. Pricing and liquidity in decentralized asset markets. *Econometrica*, Forthcoming, 2019.

Semih Üslü and Güner Velioğlu. Liquidity in the cross section of otc assets. Working paper, John Hopkins, 2019.

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

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

Sebastian Vogel. When to introduce electronic trading platforms in over-the-counter markets? Working paper, 2019.

Chaojun Wang. Core-periphery trading networks. Working paper, Wharton School, 2018.

Pierre-Olivier Weill. Essays on liquidity in financial markets, ph.d. dissertation. Stanford University, 2004.

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

Pierre-Olivier Weill. Liquidity premia in dynamic bargaining markets. *Journal of Economic Theory*, 140:66–96, 2008.

Pierre-Olivier Weill. Liquidity provision in capacity constrained markets. *Macroeconomic Dynamics*, 15(119-144), 2011.

Basil Williams. Search, liquidity, and retention: Screening multidimensional private information. Working paper, NYU, 2014.

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

Russell Wong and Mengbo Zhang. Disintermediating the federal funds market. Working paper, UCLA, 2019.

Randall Wright, Philipp Kircher, Veronica Guerrieri, and Benoit Julien. Directed search: A guided tour. NBER Working Paper 23884, 2017.

Ming Yang and Yao Zeng. The coordination of intermediation. Working paper, Duke University and University of Washington, 2019.

Abdullah Yavaş. Marketmakers versus matchmakers. *Journal of Financial Intermediation*, 2: 33–58, 1992.

Ji Hee Yoon. Endogenous market structure: Over-the-counter versus exchange trading. Working paper, 2018.

Shengxing Zhang. Liquidity missallocation in an over-the-counter market. *Journal of Economic Theory*, 174:16–56, 2017.

Haoxiang Zhu. Finding a good price in opaque over-the-counter markets. *Review of Financial Studies*, 25:1255–1285, 2011.

Junyuan Zou. Information acquisition and liquidity traps in over-the-counter markets. Working paper INSEAD, 2019.