Revisiting the Rubinstein-Wolinsky Model of Middlemen*

Grace Xun Gong
School of Economics and Academy of Financial Research, Zhejiang University

Ziqi Qiao
University of Wisconsin - Madison

Randall Wright
Zhejiang University, University of Wisconsin - Madison and NBER

November 6, 2025

Abstract

Rubinstein and Wolinsky's classic "Middlemen" paper introduced search theoretic models of intermediation. While this inspired much research, existing analysis of the model is incomplete. Rubinstein and Wolinsky show in equilibrium middlemen intermediate (buy from sellers and sell to buyers) when the rate at which they meet buyers exceeds the rate at which sellers meet buyers – but these rates should be endogenous. We characterize equilibrium in terms of fundamentals, not endogenous variables, providing novel existence, uniqueness and comparative static results. Also, being explicit about meeting technologies shows middlemen may intermediate even if their meeting technology is fundamentally inferior to sellers' technology.

Keywords: middlemen, intermediation, search

JEL classification numbers: D51, D61, D83

*We thank Boyan Jovanovic, Chao Gu, Chao He and Yu Zhu for input, as well as participants in seminars and workshops. Wright acknowledges support from the Raymond Zemon Chair in Liquid Assets and the Kenneth Burdett Professorship in Search Theory and Applications at UW. Gong acknowledges support from the Natural Science Foundation of China (No. 72503222), and the Humanities and Social Sciences Project of the Ministry of Education of China (No. 24YJC790052). The usual disclaimers apply.

1 Introduction

The classic article on "Middlemen" by Rubinstein and Wolinsky (1987), hereafter RW, introduced the search-and-bargaining approach to the study of intermediation. In terms of motivation, it is hard to beat their line:

> Despite the important role played by intermediation in most markets, it is largely ignored by the standard theoretical literature. This is because a study of intermediation requires a basic model that describes explicitly the trade frictions that give rise to the function of intermediation. But this is missing from the standard market models, where the actual process of trading is left unmodeled.

Their main result is that middlemen are active in the market, buying from sellers and selling to buyers, when they have an advantage in search, which means that middlemen are faster than sellers at contacting buyers. This may or may not be surprising, but it is certainly not something that emerges in standard, frictionless, general equilibrium theory. The RW paper has inspired much subsequent research where middlemen may have other comparative advantages: they may have lower search or storage costs; they may be able to hold larger or more diverse inventories; they may have superior information about qualitative uncertainty; they may be relatively good at bargaining; and they may be better at honoring debt obligations or enforcing the obligations of others.[1]

We revisit the original RW formulation because existing analysis of that model is incomplete. Namely, while RW show that in equilibrium middlemen play an

[1] As evidence that work following RW constitutes a vibrant research area, we could list many papers studying different ways in which middlemen may have advantages, but in the interest of space we refer to our survey, Gong et al. (2025) and the online bibliography available at https://github.com/qiao-ziqi/middlemen. Note that middlemen need not have any absolute advantage, just a comparative advantage – e.g., in Masters (2007) version of RW, middleman are not necesssarily good at search, they may just be bad at production. Also note that in addition to work following RW, there are papers using a different search-and-bargaining framework, focusing on dealers in over-the-counter asset markets following Duffie et al. (2005); see Hugonnier et al. (2025) for an extensive survey of that literature. Those models differ from RW in various ways – e.g., their dealers typically hold no inventories but simply reallocate assets across agents using a frictionless interdealer market.

active role when they have a higher arrival rate than producers in contacting consumers, these arrival rates should be *endogenous*, depending on meeting technologies as well as equilibrium behavior. We address this as follows. First, we extend RW in several ways, including heterogeneous bargaining power, which is not difficult but aids economic insight. Then we derive results nesting theirs.

Then, more significantly, rather than describing outcomes in terms of endogenous arrival rates, we characterize equilibrium trading patterns in terms of fundamentals. This generates existence and uniqueness results not in RW. It also indicates when all, some, or no middlemen are active as a function of parameters. It also clarifies some economic insights. In particular, one might think middlemen are active when the meeting technology putting them in contact with buyers is superior to the one putting sellers in contact with buyers. That is incorrect. We show that even if the technology putting them in contact with buyers is fundamentally inferior, and they have no other advantage, endogenous decisions can generate arrival rates that lead to active intermediation.

In addition, we go beyond RW by describing an explicit physical environment – let's say a micro market structure – that gives rise to bilateral meetings consistent with their reduced-form assumptions. The idea is to reinterpret RW's three-sided market, with buyers, sellers and middlemen, as segmented two-sided markets, as in more standard search models of employment (Pissarides 2000), marriage (Burdett and Coles 1997), money (Kiyotaki and Wright 1993), etc. We also provide comparative statics and some of these may be surprising – e.g., reducing frictions can lead to higher prices. Finally, we discuss how it matters whether agents exit after trading or stay in the market forever.

To summarize, our contribution is as follows:

1. We prove existence and uniqueness (not in RW).

2. We characterize equilibrium trading patterns in terms of parameters (RW only provide a relationship between trading patterns and arrival rates but arrival

rates are endogenous).

3. We develop a micro market structure consistent with their assumptions on arrival rates (RW take arrival rates as primitives);

4. We provide novel comparative static results.

The rest of the paper is organized as follows: Section 2 presents the environment and our generalized version of the main RW result. Section 3 discusses market structure in more detail. Building on that, Section 4 provides additional results and insights. Section 5 concludes.

2 Model

Time is continuous and unbounded. There are three types of agents, buyers, sellers and middlemen, that all discount the future at rate r. They meet bilaterally in a decentralized market where they trade an indivisible good, with payments made in terms of transferable utility.[2] This good is storable but at most one unit at a time. Buyers get utility u from consuming it, while sellers produce it at 0 cost merely to reduce notation. Middlemen get no utility from consuming the good and cannot produce it, but can buy it from sellers and sell it to buyers. We use the following notation: buyers and sellers are B and S; middlemen with and without a good in inventory are M_1 and M_0; and the measures of each that are active in the market at any point in time are N_b, N_s, N_1 and N_0.

As in RW, buyers and sellers flow into the market at constant rates E_b and E_s, and exit after one trade, while middlemen stay forever (but see Section 4). While inflows are exogenous, the stocks N_b and N_s are endogenous, depending on how fast they trade. The total stock of middlemen is fixed at N_m, but only a fraction

[2] While tangential to our intended contribution, it behooves us to mention our interpretation of transferable utility: a receiver of the indivisible good can produce for the provider a different good that is divisible, where p units have a disutility p from production and have a utility p from consumption. This is different from commentators (e.g., Binmore 1992) who agree that utils per se cannot be transferred, but then suggest interpreting payments as made in money. Serious work in monetary economics shows that paying with money is *not* the same as transferable utility.

τ are active – meaning those with inventory are looking to meet buyers while those without are looking to meet sellers – since for them participation has a flow cost $\kappa \geq 0$.[3] Active middlemen with (without) inventory trade whenever they meet buyers (they meet sellers), while buyers and sellers trade directly whenever they meet each other.

Let α_{ij} be the Poisson arrival rate at which type i meets j, where $i, j \in \{b, s, 1, 0\}$ indexes buyers, sellers, middlemen with inventory and middlemen without inventory. The following identities say the measure of type i meeting j is the same as the measure of type j meeting i:

$$\alpha_{bs} N_b = \alpha_{sb} N_s,\ \alpha_{b1} N_b = \alpha_{1b} N_1 \text{ and } \alpha_{s0} N_s = \alpha_{0s} N_0, \tag{1}$$

When i gives a good to j, the price p_{ij} comes from standard bargaining theory, where $\theta_{ij} \in [0, 1]$ is the share of the surplus going to type i when bargaining with type j and $\theta_{ij} = 1 - \theta_{ji}$.[4]

Let V_i be the value function for $i \in \{b, s, 1, 0\}$. Then the usual dynamic programming equations are

$$rV_b = \alpha_{bs}(u - p_{sb} - V_b) + \alpha_{b1}(u - p_{1b} - V_b) + \dot{V}_b \tag{2}$$

$$rV_s = \alpha_{sb}(p_{sb} - V_s) + \alpha_{s0}(p_{s0} - V_s) + \dot{V}_s \tag{3}$$

$$rV_1 = \alpha_{1b}(p_{1b} - V_1 + V_0) - \kappa + \dot{V}_1 \tag{4}$$

$$rV_0 = \alpha_{0s}(V_1 - V_0 - p_{s0}) - \kappa + \dot{V}_0, \tag{5}$$

where $\dot{V}_i$ is the time derivative. Notice buyers and sellers have threat points V_b and V_s but no continuation values in their surpluses as they exit after trade. In

[3] RW have $\kappa = 0$, but it is informative to generalize this even if later we let $\kappa \to 0$. Also, to be clear, buyers and sellers are always active since for them it is costless, and κ is the same for M_0 and M_1, although that can be generalized.

[4] RW impose equal bargaining powers, $\theta_{ij} = 1/2$ for all i, j, but many papers since generalize this, and it is informative, as discussed below. Also, note that with transferrable utility the outcome is the same with a variety of bargaining solutions (e.g., generalized Nash or Kalai), and i wants to trade with j if and only if j wants to trade with i if and only if the total surplus is positive.

Equation (2), e.g., the first term is the rate at which a buyer meets a seller times buyer's surplus from buyer-seller trade, and the second is the rate at which a buyer meets M_1 times buyer's surplus from buyer-middleman trade. Note that buyers' surpluses from different trades differ if $p_{sb} \neq p_{1b}$.

The prices are determined by the bargaining outcomes. Recall the surplus for i from ij trade, denoted by Σ_{ij}, are

$$\Sigma_{bs} = u - p_{sb} - V_b \text{ and } \Sigma_{sb} = p_{sb} - V_s$$
$$\Sigma_{b1} = u - p_{1b} - V_b \text{ and } \Sigma_{1b} = p_{1b} - V_1 + V_0$$
$$\Sigma_{0s} = V_1 - V_0 - p_{s0} \text{ and } \Sigma_{s0} = p_{s0} - V_s.$$

Since i's surplus share from ij trade is θ_{ij}, we have $\Sigma_{ij} = \theta_{ij}(\Sigma_{ij} + \Sigma_{ji})$. It follows that in terms of the value functions, the bargained prices are

$$p_{sb} = \theta_{sb}(u - V_b) + \theta_{bs} V_s \tag{6}$$
$$p_{1b} = \theta_{1b}(u - V_b) + \theta_{b1}(V_1 - V_0) \tag{7}$$
$$p_{s0} = \theta_{s0}(V_1 - V_0) + \theta_{0s} V_s, \tag{8}$$

which are weighted averages of the costs and benefits from ij.

The best response condition for middlemen's participation is

$$\tau \begin{cases} = 1 & \text{if } V_0 > 0 \\ \in [0,1] & \text{if } V_0 = 0 \\ = 0 & \text{if } V_0 < 0 \end{cases} \tag{9}$$

From the perspective of an inactive middleman, $V_0 > 0$ implies becoming M_0 generates positive return, while $V_0 < 0$ implies the opposite. Note that V_1 affects participation indirectly, which is captured by Equation (5).

In addition, the laws of motion for the state variables are

$$\dot{N}_b = E_b - (\alpha_{bs} + \alpha_{b1}) N_b \tag{10}$$
$$\dot{N}_s = E_s - (\alpha_{sb} + \alpha_{s0}) N_s \tag{11}$$
$$\dot{N}_1 = \alpha_{0s} N_0 - \alpha_{1b} N_1, \tag{12}$$

where E_i is the inflow of type $i \in \{B, S\}$, with the identity $\tau N_m = N_1 + N_0$ and N_m is the measure of all (active and inactive) middlemen. In Equation (10), e.g., the first term is the inflow of buyers, E_b, and the second the outflow, those buying from sellers at rate α_{bs} plus those buying from M_1 at rate α_{b1}.

This environment is identical to RW if $\theta_{ij} = 1/2$ and $\kappa = 0$. Furthermore, we adopt their restrictions on meeting technologies without question, for now, and discuss microfoundations later. To proceed, notice there are three types of trade that we call: *direct* trade (D) between buyers and sellers; *wholesale* trade (W) between sellers and M_0; and *retail* trade (R) between M_1 and buyers. Let $\mu_i : \mathbb{R}_+^2 \to \mathbb{R}_+$ be the meeting function, assumed strictly increasing. For each type: the measure of direct trade is $\mu_D(N_b, N_s)$; the measure of wholesale trade is $\mu_W(N_0, N_s)$; and the measure of retail trade is $\mu_R(N_b, N_1)$. As in textbook search theory, the α's satisfy $\alpha_{bs} = \mu_D(N_b, N_s)/N_b$, $\alpha_{sb} = \mu_D(N_b, N_s)/N_s$, etc., and if the μ's display constant returns to scale, α_{ij} depends only on the ratio N_i/N_j, where the buyer-seller ratio is called market tightness.[5]

As in RW, we make the wholesale and retail meeting functions the same:

Assumption 1 $\mu_W(\mathbf{n}) = \mu_R(\mathbf{n})$ *for all* $\mathbf{n} \in \mathbb{R}_+^2$.

While this eases the presentation, Propositions 2 and 4 below do not actually use Assumption 1, and while it is sufficient for Proposition 1 it is not necessary.

Also as in RW, we focus on steady state, where $\dot{N}_i = \dot{V}_j = 0$, and on outcomes that are symmetric in the sense that $N_s = N_b$, which requires $E_s = E_b \equiv E$.

Lemma 1 *Under Assumption 1 and* $E_s = E_b$*, in symmetric steady state* $\alpha_{bs} = \alpha_{sb}$*,* $\alpha_{b1} = \alpha_{s0}$*,* $\alpha_{1b} = \alpha_{0s}$*, and* $N_1 = N_0$.

[5] Consider some general bilateral meeting function μ. The rate α_{ij} at which i meets j is the total meetings $\mu(N_i, N_j)$ between i and j divided by N_i the number of i agents. If μ displays constant returns to scale, then $\alpha_{ij} = \mu(N_i, N_j)/N_i = \mu(1, N_j/N_i)$, which depends only on N_j/N_i or equivalently N_i/N_j.

Proof: In steady state $N_1\alpha_{1b} = N_0\alpha_{0s}$. By Equation (1), $N_1\alpha_{1b} = N_b\alpha_{b1}$ and $N_0\alpha_{0s} = N_s\alpha_{s0}$. Hence $N_b\alpha_{b1} = N_s\alpha_{s0}$. When $N_s = N_b$, $\alpha_{b1} = \alpha_{s0}$, and $N_0 = N_1$ by Assumption 1. ■

Lemma 1 establishes the tractability of the symmetric steady state due to the pairwise-equal α's and N's. This result is intuitive: in a steady state, equal entry rates for sellers and buyers necessitate equal exit rates. Since the meeting technologies are symmetric between sellers and buyers, the α's and the N's must also be symmetric.

Now use Equations (6)-(8) to eliminate p's and impose steady state to reduce Equations (2)-(5) to

$$rV_b = \alpha_{bs}\theta_{bs}(u - V_b - V_s) + \alpha_{b1}\theta_{b1}(u - V_b - V_1 + V_0) \tag{13}$$

$$rV_s = \alpha_{sb}\theta_{sb}(u - V_b - V_s) + \alpha_{s0}\theta_{s0}(V_1 - V_0 - V_s) \tag{14}$$

$$rV_1 = \alpha_{1b}\theta_{1b}(u - V_b - V_1 + V_0) - \kappa \tag{15}$$

$$rV_0 = \alpha_{0s}\theta_{0s}(V_1 - V_0 - V_s) - \kappa. \tag{16}$$

Then use Lemma 1 and Equations (10)-(12) to derive the steady state conditions

$$E = (\alpha_{bs} + \alpha_{b1})N \tag{17}$$

$$\tau N_m = 2N_1 = 2N_0, \tag{18}$$

where $N_s = N_b \equiv N$. In Equation (17), the term on the left is the inflow of buyers, and the term on the right is the outflow of buyers who trade with sellers at rate α_{bs} and with M_1 at rate α_{b1}. The seller's steady date condition is same as the buyer's by Lemma 1. Then Equation (18) simply uses the equilibrium condition $N_1 = N_0$.

Let $\mathbf{V}$, $\mathbf{p}$ and $\mathbf{N}$ be vectors of value functions, prices and stocks. We focus on symmetric stationary equilibrium, hereafter simply *equilibrium*, defined as a list $(\mathbf{V}, \mathbf{p}, \mathbf{N}, \tau)$ such that: $\mathbf{V}$ satisfies the dynamic programming equations (13)-(16); $\mathbf{p}$ satisfies the bargaining equations (6)-(8); $\mathbf{N}$ satisfies the steady state conditions (17)-(18); and τ satisfies the best response condition (9). The proof of the following is in the Appendix:

Proposition 1 *Equilibrium exists. If the μ's display constant return to scale it is unique.*

It has been known at least since Diamond (1982) that in search models uniqueness obtains under constant return to scale, while increasing return to scale is necessary but not sufficient for multiplicity. Hence, the above result is not too surprising, but it is good to make it explicit. In fact, RW do not mention returns to scale, since they do not discuss uniqueness, or even existence; they focus only on characterizing τ in terms of the α's.

Now notice that there are three possible types of equilibria that we call regimes: no middlemen are active $\tau = 0$; all middlemen are active $\tau = 1$; and some middlemen are active $\tau \in (0, 1)$.

Proposition 2 *Define*

$$\Omega(\tau) \equiv \frac{r + \alpha_{1b}\theta_{1b} + \alpha_{0s}\theta_{0s} + \alpha_{s0}\theta_{s0}}{\alpha_{0s}\theta_{0s}(u - V_b - V_s)}. \tag{19}$$

Then equilibrium with $\tau = 0$ exists if and only if $\alpha_{1b}\theta_{1b} \le \alpha_{sb}\theta_{sb} + \kappa\Omega(0)$; equilibrium with $\tau = 1$ exists if and only if $\alpha_{1b}\theta_{1b} \ge \alpha_{sb}\theta_{sb} + \kappa\Omega(1)$; equilibrium with $\tau \in (0, 1)$ exists if and only if $\alpha_{1b}\theta_{1b} = \alpha_{sb}\theta_{sb} + \kappa\Omega(\tau)$.

Proof: The surplus from wholesale trade satisfies

$$V_1 - V_0 - V_s = \frac{(u - V_b - V_s)(\alpha_{1b}\theta_{1b} - \alpha_{sb}\theta_{sb})}{r + \alpha_{1b}\theta_{1b} + \alpha_{0s}\theta_{0s} + \alpha_{s0}\theta_{s0}} = \frac{\alpha_{1b}\theta_{1b} - \alpha_{sb}\theta_{sb}}{\alpha_{0s}\theta_{0s}\Omega(\tau)}.$$

By Equation (9), $\tau = 0$ requires $\alpha_{0s}\theta_{0s}(V_1 - V_0 - V_s) \le \kappa$. Hence $\tau = 0$ is consistent with equilibrium if $\alpha_{1b}\theta_{1b} \le \alpha_{sb}\theta_{sb} + \kappa\Omega(0)$. Similar logic applies to the other regimes. ■

Remark 1 *If the α's and Ω were constants this would partition parameter space into regions where each regime is an equilibrium; but they are not constants.*

Proposition 2 establishes the equilibrium condition for middlemen's participation. This condition consists of two components. The first component is the term $\alpha_{1b}\theta_{1b} - \alpha_{sb}\theta_{sb}$, which compares the relative advantage of the middleman versus the seller in trading with a buyer. The second component is the middleman's weighted flow cost. The weight increases with the discount rate but decreases with the gains from a complete seller-to-buyer trade.

Corollary 1 *Suppose $\kappa = 0$. Then equilibrium with $\tau = 0$ exists if and only if $\alpha_{1b}\theta_{1b} \leq \alpha_{sb}\theta_{sb}$; equilibrium with $\tau = 1$ exists if and only if $\alpha_{1b}\theta_{1b} \geq \alpha_{sb}\theta_{sb}$; equilibrium with $\tau \in (0,1)$ exists if and only if $\alpha_{1b}\theta_{1b} = \alpha_{sb}\theta_{sb}$.*

Remark 2 *The RW result is a special case of Corollary 1 under their assumption $\theta_{ij} = 1/2$ for all i, j, although it really only requires $\theta_{1b} = \theta_{sb}$.*

Intuitively, the RW result says that middlemen have a role when α_{1b} exceeds α_{sb}, as stated in Remark 2 with $\theta_{1b} = \theta_{sb}$. More generally, Corollary 1 shows the arrival rates α_{1b} and α_{sb} must be adjusted for bargaining powers θ_{1b} and θ_{sb}.[6] Still more generally, Proposition 2 gives a further adjustment for $\kappa > 0$. As intuitive as these results may be, note that the original RW paper takes $\Omega(\tau)$ and the α's as given, and a main point here is to make them endogenous.

To be clear, the RW result does *not* say that middlemen are active when they contact buyers via a meeting technology that is superior to the one via which sellers contact buyers. We show below that when the meeting technology of middlemen, μ_R, is fundamentally inferior to that of sellers, μ_D, middlemen can be active even if $\theta_{1b} = \theta_{sb}$ and $\kappa = 0$, and sometimes even if $\theta_{1b} > \theta_{sb}$ and $\kappa > 0$. Intuitively, middlemen may face an inferior meeting technology, suggesting τ should be 0; but if τ is small, their arrival rate α_{1b} can be big even

[6] One might ask why these results depend on θ_{1b} and θ_{sb} but not on θ_{0s}. Intuitively, the reason is that θ_{0s} affects how the surplus is split in trades between middlemen and sellers but does not affect whether they trade – given transferable utility, they trade whenever the joint surplus is positive.

with an inferior μ_R. Without claiming this is "deep" we contend that it leads to interesting insights.

3 Meeting Technologies

At this point it is useful to delve into how the meeting process operates. While the RW specification is in some ways general, it is in other ways quite special since it restricts meetings between i and j to depend only on the mass of i and j, which is violated by many standard models.[7] Indeed, it is violated in middleman models that use *uniform random matching*, which means that the probability i meets j is proportional to the fraction of type j in the market – and therefore, e.g., the rate at which a seller meets a buyer depends on all the N's, not just N_s and N_b.

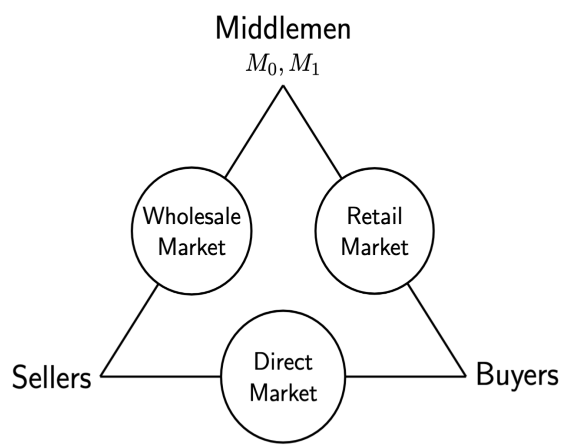

Figure 1: Market Structure

We propose a market structure based on spatial separation that provides an interpretation of RW. As shown in Figure 1, different types are located at distinct points represented by nodes on a triangle. There are three submarkets along the

[7]Consider a labor market with $\mu(N_v, N_u)$, where N_u and N_v measure unemployment and vacancies. Some papers (e.g., Albrecht and Vroman 2002) have high- and low-skill workers. Even if a firm only wants high-skill workers, they might meet low-skill workers, so the presence of low-skill workers affects the firm's arrival rate of high-skill workers. That is inconsistent with the RW specification. The situation is similar in goods markets (e.g., see Bethune et al. 2020).

edges of the triangle: a direct market with buyers and sellers; a wholesale market with sellers and active middlemen without inventory; and a retail market with buyers and active middlemen with inventory. For middlemen search is directed – those with inventory are only in retail market, those without are only in wholesale market. For buyers and sellers, everyone goes to the two closest markets, but not the third – it's just too far.[8]

Assume buyers can participate in both the direct and retail markets at the same time, and sellers can participate in both the direct and wholesale markets at the same time. One interpretation invokes the two-person households used in cash-in-advance models (e.g., Lucas 1980), even if they typically have one shopper and one worker, while here it would be two shoppers for buyer households and two workers for seller households. Another interpretation is telephone matching (e.g., Mortensen and Pissarides 1999), where our agents post their phone numbers on bulletin boards in the two nearby markets, but not the distant market, maybe again because it's too far, or maybe now because long-distant calls are too expensive. In any case, while agents participate in two submarkets, the probability of two arrivals at any point in time is 0 given independent Poisson arrivals.[9]

This generates a meeting process consistent with RW. One might say it replaces a three-sided market by three two-sided markets (similar to Gong and Wright 2024, except there temporal, not spatial, separation is at work). In particular, direct meetings now depend on (N_b, N_s) and not other N's, wholesale meetings depend on (N_0, N_s) and not other N's, and retail meetings depend on

[8] Figure 1 is reminiscent of Kiyotaki and Wright (1989), which was designed to think about money, but can be interpreted in terms of intermediation: When agent i at one node acquires good j from someone at another node, then later trades it to someone at a third node, we can say good j serves as commodity money, but we can also say agent i serves as a middleman. Analysis in that framework usually uses uniform random matching, however, while the way we use it below has a flavor of (partially) directed search, similar to Corbae et al. (2003).

[9] Note that it is exogenous for now that one member of a household goes, or one telephone number is posted, in each nearby market, and not both in the same nearby market, but one could presumably endogenize that. This is indicative of the notion that it is always possible to pursue further microfoundations for microfoundations.

(N_b, N_1) and not other N's consistent with RW. With a clearer understanding of this we can say much more about the RW model.

4 Beyond RW

Here we let $\kappa \to 0$ and suppose the meeting function in each submarket is a constant times the function $\mu(\cdot)$.[10] Assumption 1 implies the constants must be the same in the wholesale and retail markets, but the constant in the direct market can be different. Hence,

$$\mu_D = \delta\mu(N_b, N_s),\ \mu_W = \sigma\mu(N_0, N_s) \text{ and } \mu_R = \sigma\mu(N_b, N_1). \tag{20}$$

where δ and σ represent *fundamental* properties – i.e., exogenous parameters – of the meeting technologies.

We now reduce the equilibrium to two equations in (N, τ), with details in the Appendix. From Equation (17), we get $N = N_\tau$, a function of τ. Then define $\Phi : [0, 1] \to \mathbb{R}_+$ by

$$\Phi(\tau) \equiv \frac{\tau N_m \mu(1,1)}{\mu(2N_\tau, \tau N_m)}.$$

Since $\Phi'(\tau) > 0$, it is invertible, and we can write:

$$\tau = \begin{cases} 1 & \text{if } \sigma\theta_{1b} > \delta\theta_{sb}\Phi(1) \\ \Phi^{-1}\left(\frac{\sigma\theta_{1b}}{\delta\theta_{sb}}\right) & \text{if } \delta\theta_{sb}\Phi(0) \le \sigma\theta_{1b} \le \delta\theta_{sb}\Phi(1) \\ 0 & \text{if } \sigma\theta_{1b} < \delta\theta_{sb}\Phi(0) \end{cases}$$

This partitions parameter space into three regions where each regime constitutes the unique equilibrium.

The left panel of Figure 2 shows the result in (σ, θ_{1b}) space. In terms of economics, when σ and θ_{1b} are low middlemen are inactive, naturally, as they are bad at both search and bargaining. When σ and θ_{1b} are higher, some middlemen will be active, but not all of them because that would make their arrival rate too low to satisfy the best response condition. When σ and θ_{1b} are higher still

[10] The unique existence directly follows Proposition 1 and does not rely on these assumptions.

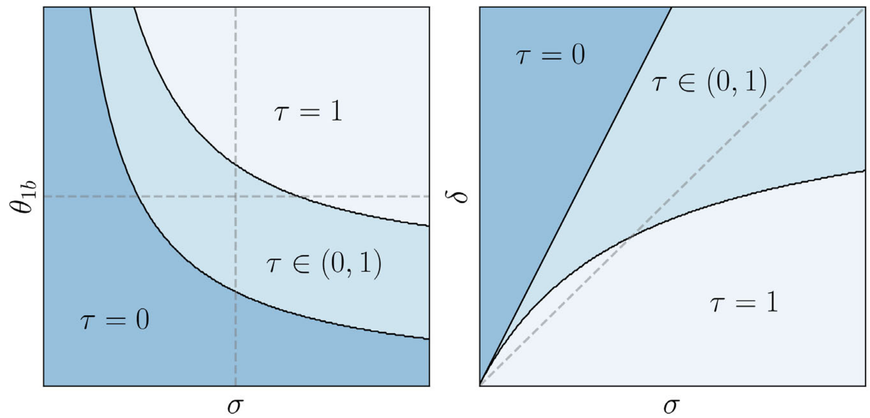

Figure 2: Equilibrium Regimes

all middlemen will be active. The two dashed lines show sellers' corresponding parameters, $\theta_{1b} = \theta_{sb}$ and $\sigma = \delta$. Notice $\tau > 0$ when middlemen and sellers are the same in terms of their meeting technologies and bargaining powers.

Proposition 3 *Given $\delta\theta_{sb} > 0$, $\tau > 0$ is the equilibrium for $\sigma\theta_{1b} = \delta\theta_{sb}$ and for some $\sigma\theta_{1b} < \delta\theta_{sb}$. Moreover, $\tau = 1$ is the equilibrium for such $\sigma\theta_{1b}$ if E/N_m is big.*

Proof: Because $\Phi(0) < 1$, $\delta\theta_{sb}\Phi(0) < \delta\theta_{sb}$. So $\tau > 0$ is an equilibrium if $\delta\theta_{sb}\Phi(0) < \sigma\theta_{1b} < \delta\theta_{sb}$. And, as $E/N_m \to \infty$, $\Phi(1) \to \Phi(0)$. Hence $\tau = 1$ is an equilibrium if $\sigma\theta_{1b} > \delta\theta_{sb}\Phi(0)$. ■

Proposition 3 formalizes what Figure 2 illustrates. Specifically, middlemen always participate if they possess the same meeting technology and bargaining power as sellers. Then by continuity, they also participate even when their technologies are somewhat inferior. In such cases, all middlemen would be active if the aggregate inflow of buyers and sellers is significantly larger than the measure of middlemen. The intuition is that, under these conditions, active middlemen enjoy a sufficiently high arrival rate in the retail market because the market tightness highly favors them, thus compensating for their inferior meeting technology or bargaining power.

RW focus on α's. For a direct comparison, the right panel of Figure 2 fixes $\theta_{1b} = \theta_{sb} = 1/2$ and compares the meeting technologies. Given the same δ, τ increases in σ; given the same σ, τ decreases in δ. When $\sigma = \delta$, we always have $\tau > 0$. The intuition is simple: if $\tau = 0$, then $N_b/N_1 = \infty$, and given the same fundamental technologies $\alpha_{1b} > \alpha_{sb}$ for sure.[11] This again highlights the importance of accounting for market tightness in equilibrium analysis. Moreover, when $\sigma = \delta$, τ is high when both have a poor meeting technology and low when both have a good technology. This is because better meeting technology leads to more trades, and consequently leads to lower N_b and N_s as buyers and sellers leave after trade, so an overall improvement in the meeting technology implies lower arrival rates of counterparties for active middlemen in retail and wholesale market, and consequently, lower τ. That would not be apparent if one looked only at the RW result.

Figure 3 displays some comparative statics.[12] On the left, panel (a) varies σ with $\theta_{1b} = \theta_{0s} = 0.9$; on the right, panel (b) varies $\theta_{1b} = \theta_{0s}$ with $\sigma = 0.5$. The top row shows (N, τ), where it is clear that differently shaded regions indicate different regimes, while the bottom row shows prices. Higher σ or $\theta_{1b} = \theta_{0s}$ increases τ and decreases N, but increases in bargaining power only strictly decrease N when $\tau < 1$, while increases in search efficiency always strictly decrease N. For prices, the dotted segments show potential (off the equilibrium path) prices for middlemen when $\tau = 0$. All prices decrease with σ since two forces operate in the same direction: higher τ increases competition and decreases N. Prices are non-monotone in $\theta_{1b} = \theta_{0s}$ because competition dominates when τ is low, so

[11] An anonymous commentator offered the following intuition. Suppose there is a large and equal number of buyers and sellers and suppose further that there are few middlemen. Most trade will be between buyers and sellers, but it is still profitable for a middleman to be active because the meeting rate depends on the buyer-middleman ratio (or the seller-middleman ratio in the other submarket): the fewer middlemen there are, the higher the meeting rate for middlemen.

[12] This is an example using uniform random matching in each submarket, $E = 1$, $N_m = 5$, $\theta_{sb} = \delta = 0.5$, $u = 1$, $\kappa = 0$ and $r = 0.03$. The results are not sensitive to these parameters, except for κ, which cannot be excessively large or middleman would not participate.

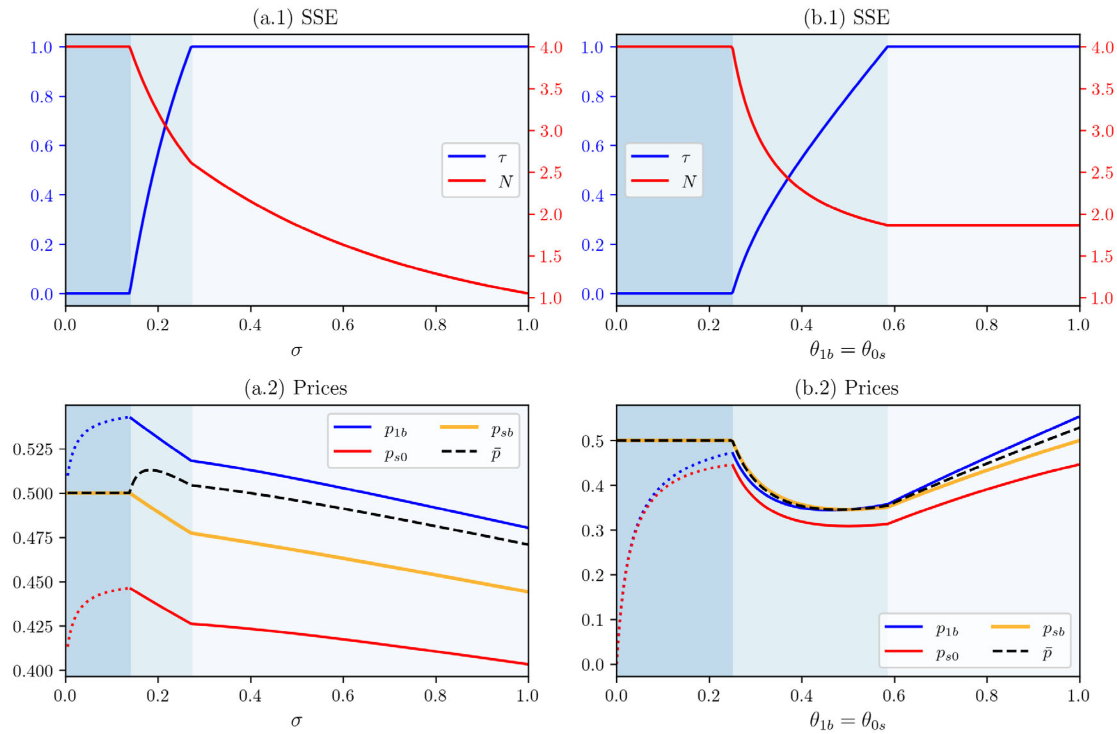

Figure 3: Parameter Effects

prices fall, but when τ is big the impact of a marginal participant on competition diminishes and bargaining power dominates.

The average price paid by buyers, denoted $\bar{p}$, is non-monotonic in σ due to a composition effect: both p_{sb} and p_{1b} fall with σ, but since higher τ increases the size of retail trade relative to direct trade, and the retail price is above the direct price, $\bar{p}$ can increase.[13] One can also check that price dispersion measured by the coefficient of variation can be non-monotone, first increasing then decreasing in σ. These results are interesting in light of some commentary – e.g., Ellison and Ellison (2005) say "evidence from the Internet... challenged the existing search models, because we did not see the tremendous decrease in prices and price dispersion that many had predicted," while Baye et al. (2006) say "Reductions

[13] The composition effect is contingent upon a high θ_{1b}. Intuitively, for middleman entry to result in a higher average price, they must possess substantial pricing power. From another perspective, a superior meeting technology does not necessarily lead to lower market prices if it is coupled with a sufficiently high ability to set prices.

in information costs over the past century have neither reduced nor eliminated the levels of price dispersion observed." This example demonstrates how search theory does *not* predict average price or price dispersion must fall with reductions in frictions, which is one good reason for pushing RW further than previous studies.[14]

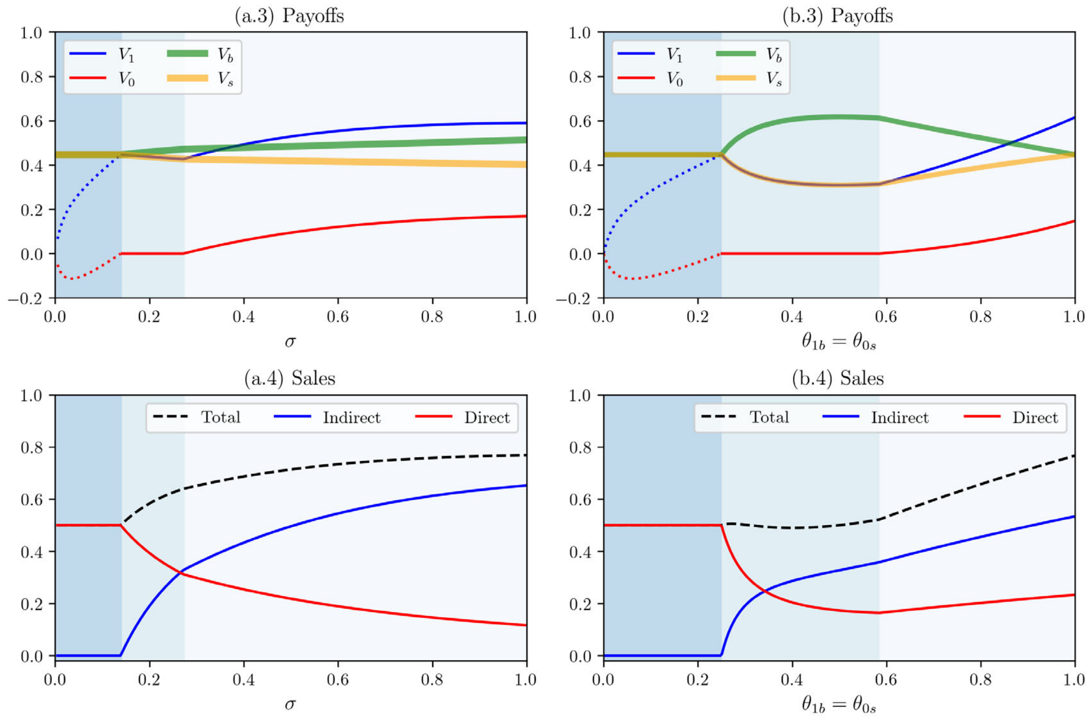

Figure 4: More Parameter Effects

Figure 4 shows the impact of search efficiency and bargaining power on payoffs and sales, where again dotted segments show potential (off the equilibrium path) values for middlemen when $\tau = 0$. Notice V_b and V_s are monotone in σ and non-monotone in $\theta_{1b} = \theta_{0s}$. This issue has been discussed extensively elsewhere, and, in general, it is known that middlemen can either increase or decrease welfare

[14] Having price dispersion non-monotone in frictions should be unsurprising to those who know search theory since then well known model of Burdett and Judd (1983) gets that. Having the average price non-monotone is harder. Lester (2011) gets that in a finite-agent model, due to strategic considerations, while Bethune et al. (2020) get it in a monetary model, where lower frictions mean buyers carry more cash, so sellers can charge more; those effects are not in play here.

depending on the specific model details.[15]

We have one last point. RW impose many symmetry assumptions on primitives, like $E_s = E_b$, $\theta_{ij} = 1/2$ and $\mu_W = \mu_R$, and focus on symmetric outcomes with $N_s = N_b$, but there is one stark asymmetry: middlemen stay in the market forever while others exit after one trade. For buyers, that does not really matter, but the asymmetry between sellers and middlemen might. Instead of RW's specification, with long-lived middlemen and short-lived sellers, consider the case where both exit after one sale. It is not hard to show by emulating the above methods that when $\kappa = 0$ our version of RW (Corollary 1) holds exactly as written.

Now consider imposing symmetry by letting sellers stay in the market forever, producing another good immediately after each sale. Then $\tau = 1$ is always consistent with the equilibrium, provided the middlemen meeting technology is not too inefficient and bargaining power is not too low. The reason is that now sellers have no opportunity cost to wholesale trade, since upon giving middlemen a good sellers get another one. So they trade as long as parameters for middlemen are not too unfavorable – e.g., θ_{1b} cannot be too low or middlemen cannot recover the cost of wholesale trade due to the usual holdup problem, given the payment to sellers is sunk when middlemen contact buyers.[16] In any event, this version demonstrates plainly how middlemen can be fundamentally inferior to sellers yet still active in equilibrium.

[15] The efficiency properties equilibrium with middlemen in general depend on many factors, including the potential multiplicity of equilibria (Nosal et al. 2019), occupational choice (Li 1998), private information (Biglaiser and Li 2018), market structure (Awaya et al. 2025), inventory technologies (Shevchenko 2004), and the pricing mechanism (Watanabe 2010). See the Gong et al. (2025) survey for a discussion of the framework including these elements.

[16] RW discuss how consignment sales (middlemen pay sellers only after trading with buyers) avoid this holdup problem. Since this is well understood we do not further pursue it here. What may be more interesting is to consider price posting and directed search, rather than bargaining and random search, which is another way to avoid holdup problems (Wright et al. 2001). Note that Watanabe (2010,2020) and Gautier et al. (2023) already discuss middlemen with posting and directed search, but there is more to be done along these lines.

5 Conclusion

We revisited a canonical search-and-bargaining model of middlemen. While RW made a big contribution by introducing the framework, there were loose ends. Their result that middlemen are active when they meet buyers faster than sellers meet buyers is interesting, and extended versions can accommodate heterogeneous bargaining powers, participation costs, etc., but it is a characterization of an equilibrium outcome – middlemen activity – in terms of another equilibrium outcome – arrival rates. We characterized middlemen activity and arrival rates in terms of fundamentals, providing existence, uniqueness and comparative static results not in previous papers.

We also discussed how details matter, such as whether agents are in the market for a short or long time. Perhaps most significantly, delving deeper into meeting technologies we clarified this: it is not the case that middlemen have a role when the technology connecting them to buyers is superior to the one connecting sellers to buyers. Even if the former technology is fundamentally inferior, middlemen can be active, given the way equilibrium arrival rates adjust to their activity.

RW was published some time ago, but is still relevant and continues to influence good research. However, it seems fair to suggest that search-based theories of intermediation have not had as big an impact as similar models of labor, marriage, housing, etc. Perhaps one reason it that the baseline RW framework was not totally transparent and had not been analyzed completely. Our goal was to rectify this and develop additional insights along the way. The findings have helped us better understand this model and better understand search theory more generally.

Appendix

Proof of Proposition 1: (Existence) The dynamic programming equations are linear in $\mathbf{V}$ with a unique solution. Write $V_0 = g(N_\tau, \tau)$, where N_τ is the solution to $E = (\alpha_{bs} + \alpha_{b1})N_\tau$, which exists if $\lim_{N\to\infty} \mu_D(N, N) > E$. Then, if $g(N_\tau, \tau) = 0$ has a solution for $\tau \in [0, 1]$, it is an equilibrium. Otherwise, $\tau = 1$ if $g(\cdot) > 0$ or $\tau = 0$ if $g(\cdot) < 0$ is an equilibrium.

(Uniqueness) Constant return to scale of $\mu_D(\cdot)$ implies $\alpha_{bs} = \alpha_{sb} = \mu_D(1, 1)$ is constant in any equilibrium. Constant return to scale in $\mu_W(\cdot) = \mu_R(\cdot)$ (they are the same by Assumption 1) implies $\alpha_{1b,\tau=0} = \lim_{A\to 0} \mu_R(N_{\tau=0}/A, 1) > \mu_R(2N_{\tau=1}/N_m, 1) = \alpha_{1b,\tau=1}$. Whenever $\alpha_{1b,\tau=0}\theta_{1b} \leq \alpha_{sb}\theta_{sb} + \kappa\Omega(0)$, $\alpha_{1b,\tau=1}\theta_{1b} \leq \alpha_{sb}\theta_{sb} + \kappa\Omega(0)$; whenever $\alpha_{1b,\tau=1}\theta_{1b} \geq \alpha_{sb}\theta_{sb} + \kappa\Omega(1)$, $\alpha_{1b,\tau=0}\theta_{1b} \geq \alpha_{sb}\theta_{sb} + \kappa\Omega(1)$. Given Proposition 2, the uniqueness can be established by $\Omega(0) < \Omega(1)$ where

$$\begin{aligned}\Omega(\tau) = &\frac{(\theta_{0s} + \theta_{1b})(r + \alpha_{bs} + \alpha_{sb}) + \alpha_{b1}\theta_{b1}\theta_{0s} + \alpha_{s0}\theta_{s0}\theta_{1b}}{ru\theta_{0s}} \\ &+ \frac{r(r + \alpha_{bs} + \alpha_{sb} + \alpha_{b1}\theta_{b1}) + \alpha_{s0}\theta_{s0}(r + \alpha_{bs}\theta_{bs} + \alpha_{b1}\theta_{b1}) + \alpha_{sb}\theta_{sb}\alpha_{b1}\theta_{b1}}{ru\alpha_{0s}\theta_{0s}}\end{aligned}$$

Recall α_{bs}, α_{sb} and θ's are constant. By Lemma 1, $\alpha_{b1} = \alpha_{s0}$ increases in τ, and $\alpha_{1b} = \alpha_{0s}$ decreases in τ. The numerators are increasing in τ, while the denominators are non-decreasing. Therefore, $\Omega'(\tau) > 0$ and $\Omega(0) < \Omega(1)$. ■

Details for Section 4: Let $\bar{\mu} \equiv \lim_{N_0\to 0} \mu(N/N_0, 1)$ be the arrival rate for M_0 in wholesale market when $\tau = 0$, and $\hat{\mu} \equiv \mu(1, 1)$ the arrival rate for buyers and sellers in the direct market. Since $\mu(\cdot)$ is strictly increasing, $\bar{\mu} > \hat{\mu}$. From Equation (18),

$$\frac{E}{N_m} = \delta\hat{\mu}\frac{N}{N_m} + \frac{\sigma}{2}\mu\left(\frac{2N}{N_m}, \tau\right).$$

The LHS is constant, while the RHS is strictly increasing in N and τ. Hence, there exists $f : [0, 1]\times\mathbb{R}_+ \to \mathbb{R}_+$ with $f_1 < 0 < f_2$ such that $N_\tau = N_m f(\tau, E/N_m)$.

Substitute N_τ into the best response condition and define $\Phi : [0,1] \to \mathbb{R}_+$ as

$$\Phi(\tau) \equiv \frac{\hat{\mu}}{\mu\left(2N_\tau/\tau N_n, 1\right)} = \frac{\hat{\mu}}{\mu\left(2f(\tau, E/N_m)/\tau, 1\right)}.$$

Note that $\Phi(0) = \hat{\mu}/\bar{\mu} < 1$. Since $f_1 < 0$ and $\mu_1 > 0$, we have $\Phi'(\tau) > 0$. Then we have:

$$\tau = \begin{cases} 1 & \text{if } \sigma\theta_{1b} > \delta\theta_{sb}\Phi(1) \\ \Phi^{-1}\left(\frac{\sigma\theta_{1b}}{\delta\theta_{sb}}\right) & \text{if } \delta\theta_{sb}\Phi(0) \leq \sigma\theta_{1b} \leq \delta\theta_{sb}\Phi(1) \\ 0 & \text{if } \sigma\theta_{1b} < \delta\theta_{sb}\Phi(0) \end{cases}$$

Consider the three regimes in (θ_{1b}, σ) space. There are two cutoffs. The first separates $\tau = 0$ and $\tau > 0$, and it depends only on primitives. The second cutoff separates $\tau < 1$ and $\tau = 1$, and can be represented by

$$\sigma\theta_{1b} = \delta\theta_{sb}\Phi(1) = \frac{\delta\theta_{sb}\hat{\mu}}{\mu\left(2N_{\tau=1}/N_m, 1\right)}.$$

As $\Phi(1) > \hat{\mu}/\bar{\mu} = \Phi(0)$, the second cutoff is above the first, confirming uniqueness. Observe that $\sigma\mu\left(2N_{\tau=1}/N_m, 1\right)$ is increasing in σ. Therefore, θ_{1b} decreases in σ at the cutoff. Note that the curvature is ambiguous, depending on the meeting function; however, for popular choices of μ, like uniform, urn-ball or Cobb-Douglas, the second cutoff is convex.

References

J. Albrecht and S. Vroman (2002) "A Matching Model with Endogenous Skill Requirements," *International Economic Rreview* 43, 283-305.

Y. Awaya, K. Iwasaki and M. Watanabe (2025) "Money is the root of all bubbles," mimeo.

M. Baye, J. Morgan and P. Scholten (2006) "Information, search, and price dispersion," *Handbook of Economics and Information Systems*, Elsevier, 323-75.

K. Binmore (1992) *Fun and Games: A Text on Game Theory*, D.C. Heath.

Z. Bethune, M. Choi and R. Wright (2020) "Frictional Goods Markets: Theory and Applications," *Review of Economic Studies* 87, 691-720.

G. Biglaiser and F. Li (2018) "Middlemen: The good, the bad, and the ugly," *RAND Journal of Economics* 49, 3-22.

K. Burdett, and M. Coles (1997) "Marriage and Class," *Quarterly Journal of Economics* 112, 141-68.

K. Burdett and K. Judd (1983) "Equilibrium Price Dispersion," *Econometrica* 51, 955–69.

D. Corbae,T. Temzelides and R. Wright (2003) "Directed Matching and Monetary Exchange," *Econometrica* 71, 731-56.

P. Diamond (1982) "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy* 90, 881-94.

D. Duffie, N. Garleanu and L. Pederson (2005) "Over-the-Counter Markets," *Econometrica* 73, 1815-47.

G. Ellison and S. Ellison (2005) "Lessons About Markets from the Internet," *Journal of Economic Perspectives* 19, 139-58.

P. Gautier, B. Hu and M. Watanabe (2023) "Marketmaking Middlemen," *RAND Journal of Economics* 54, 83-103.

G. Gong, Z. Qiao and R. Wright (2025) "Middlemen in Search Equilibrium: A Survey, " *Elements in Money and Banking*, edited by Chao Gu and Joseph Haslag, Cambridge University Press, in print.

G. Gong and R. Wright (2024) "Middlemen in Search Equilibrium with Intensive and Extensive Margins," *International Economic Review* 65, 1657-79.

J. Hugonnier, B. Lester and P. Weill (2025) *The Economics of Over-the-Counter Markets*, Princeton.

N. Kiyotaki and R. Wright (1989) "On Money as a Medium of Exchange," *Journal of Political Economy* 97, 927-54.

N. Kiyotaki and R. Wright (1993) "A Search Theoretic Model of Monetary Economics," *American Economic Review* 83, 63-77.

B. Lester (2011) "Information and Prices with Capacity Constraints," *American Economic Review* 101, 1591-600.

Y. Li (1998) "Middlemen and private Information," *Journal of Monetary Economics* 42, 131-159.

R. Lucas (1980) "Equilibrium in a Pure Currency Economy," *Economic Inquiry* 18, 203-20.

A. Masters (2007) "Middlemen in Search Equilibrium," *International Economic Review* 48, 343-362.

D. Mortensen and C. Pissarides (1999) "New Developments in Models of Search in the Labor Market," *Handbook of Labor Economics* 3, 2567-627.

E. Nosal, Y. Wong and R. Wright (2019) " Intermediation in Markets for Goods and Markets for Assets," *Journal of Economic Theory* 183, 876-906.

C. Pissarides (2000) *Equilibrium Unemployment Theory*, MIT Press.

A. Rubinstein and A. Wolinsky (1987) "Middlemen," *Quarterly Journal of Economics* 102, 581-94.

A. Shevchenko (2004) "Middlemen," *International Economic Review* 45, 1-24.

M. Watanabe (2010) "A Model of Merchants," *Journal of Economic Theory* 145, 1865-89.

M. Watanabe (2020) "Middlemen: A Directed Search Equilibrium Approach," *BE Journal of Macroeconomics*

R. Wright. P. Kircher, B. Julien and V. Guerrieri (2021) "Directed Search and Competitive Search Equilibrium: A Guided Tour" *Journal of Economic Literature* 59, 90-148.