Middlemen Redux*

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

March 17, 2025

Abstract

Rubinstein and Wolinsky's "Middlemen" introduced search models of intermediation. While this inspired much research, existing analysis of their model is incomplete. They 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. However, these rates depend on agents' decisions. We characterize equilibrium in terms of fundamentals, not endogenous variables, providing existence, uniqueness and (some surprising) comparative static results not in previous work. Also, being explicit about meeting technologies shows middlemen may intermediate even if their technology is fundamentally inferior to sellers' technology.

Keywords: middlemen, intermediation, search, bargaining

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 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 in the model means that middlemen are faster than sellers at contacting buyers. This may or may not be surprising, but it is certainly not something seen in "standard market models" (frictionless, general equilibrium theory). In any case, RW 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]

This paper revisits the original RW formulation because their analysis is incomplete in a way that has not been addressed in the literature. Namely, while it is true that in equilibrium middlemen play an active role when they have a higher arrival rate than producers in contacting consumers, these arrival rates

[1] As evidence that research following RW constitutes a vibrant area we can list many papers studying different ways in which middlemen may have advantages. In the interest of space, this list is online at https://github.com/qiao-ziqi/middlemen, which currently includes over 50 papers with brief descriptions. We also mention that, in addition to work following RW, there are papers using a different search model, focusing on dealers in OTC asset markets, following Duffie et al. (2005); see Hugonnier et al. (2025) for a survey. Those models differ from RW in various ways – e.g., their dealers typically hold no inventories, but simply reallocate assets across investors via a frictionless interdealer market.

are *endogenous*, depending on primitive meeting technologies and on equilibrium behavior. This is addressed here as follows. First, we extend RW in several ways (e.g., heterogeneous bargaining power), which is not difficult, but aids economic insight. Then we derive results nesting theirs.

Then, more significantly, rather than describing equilibrium outcomes in terms of endogenous arrival rates, we characterize it in terms of fundamentals. This entails existence and uniqueness results not in RW, and tells us when all, some or no middlemen are active as a function of parameters. It also clarifies key economic ideas. 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 false. We show that even if the technology putting them in contact with buyers is fundamentally inferior, and they have no other advantage (e.g., their bargaining ability is no better), endogenous decisions in RW can generate arrival rates that lead to middlemen actively intermediating between buyers and sellers.

There are a few additional results. For one, we go beyond RW by describing an explicit physical environment, or market structure, that gives rise to the meeting technology in their specification, which we think may be useful in other applications.[2] We also provide various comparative statics, some of which may be surprising – e.g., it is shown how reducing search frictions can lead to higher observed prices. Finally, we discuss how it matters which agents in RW exit after trading and which stay in the market forever.

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

[2] The complication is that RW is a three-sided market with buyers, sellers and middlemen, while two-sided markets appear in standard search models of, e.g., employment (Pissarides 2000), marriage (Burdett and Coles 1997), etc.

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.[3] 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, B and S 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 τ are active – meaning those with inventory are looking to meet B while those without are looking to meet S – since for them participation has a flow cost $\kappa \geq 0$.[4] Active middlemen with (without) inventory trade whenever they meet B (they meet S), while B and S 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 with-

[3] 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. We think that studying middlemen models without transferable utility should be a priority, but is beyond the scope of the current project.

[4] RW have $\kappa = 0$, but it is informative to generalize this even if later we let $\kappa \to 0$. Also, to be clear, B and S 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.

out 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 and $\theta_{ij} = 1 - \theta_{ji}$.[5]

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 B and S have threat points V_b and V_s but no continuation values in their surpluses as they exit after trade. In (2), e.g., the first term is the rate at which B meets S times $u - p_{sb} - V_b$ and the second is the rate at which B meets M_1 times $u - p_{1b} - V_b$, which differs if $p_{sb} \neq p_{1b}$.

In terms of the value functions, bargained prices are

$$p_{sb} = \theta_{sb}(u - V_b) + \theta_{bs}V_s \tag{6}$$

$$p_{1b} = \theta_{1b}(u - V_b) + \theta_{b1}(V_1 - V_0) \tag{7}$$

$$p_{s0} = \theta_{s0}(V_1 - V_0) + \theta_{0s}V_s. \tag{8}$$

Also, the best response condition for middlemen's participation is

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

[5] 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 iff j wants to trade with i iff the total surplus is positive.

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

$$\dot{N}_b = E_b - (\alpha_{bs} + \alpha_{b1})N_b \tag{10}$$

$$\dot{N}_s = E_s - (\alpha_{sb} + \alpha_{s0})N_s \tag{11}$$

$$\dot{N}_1 = \alpha_{0s}N_0 - \alpha_{1b}N_1, \tag{12}$$

with the identity $\tau N_m = N_1 + N_0$ and N_m is the measure of all (active and inactive) middlemen. In (10), e.g., the first term is the inflow of B and the second the outflow, those buying from S plus those buying from M_1.

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 describe these, first notice there are three types of trade that we call: *direct* trade (D) between B and S; *wholesale* trade (W) between S and M_0; and *retail* trade (R) between M_1 and B. 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 (CRS), α_{ij} depends only on the ratio N_i/N_j, where the buyer-seller ratio is called market tightness.

As in RW, we make the WM and RM 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 $S = B$, which requires $E_s = E_b$.

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 $M_1 = M_0$.*

Proof: In steady state $N_1\alpha_{1b} = N_0\alpha_{0s}$. By (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. ■

Now use (6)-(8) to eliminate p's and impose steady state to reduce (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 to reduce the steady state conditions to

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

$$\tau N_m = 2A, \tag{18}$$

where $N_s = N_b \equiv N$ and $N_0 = N_1 \equiv A$ (A for *active* middlemen).

Let $\mathbf{V}$, $\mathbf{p}$ and $\mathbf{N}$ be vectors of value functions, prices and stocks. Then a symmetric stationary equilibrium (SSE) is 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 *SSE exists. If the μ's display CRS then it is unique.*

It has been known since Diamond (1982) that in this kind of model uniqueness requires CRS. RW do not mention CRS since they do not discuss uniqueness, or even existence, results that are unnecessary for their goal of characterizing τ in terms of the α's. To see what can be achieved along these lines, first note there are three possible regimes (types of equilibrium): no middlemen are active $\tau = 0$; all middlemen are active $\tau = 1$; or 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 SSE with $\tau = 0$ exists iff $\alpha_{1b}\theta_{1b} \leq \alpha_{sb}\theta_{sb} + \kappa\Omega(0)$; SSE with $\tau = 1$ exists iff $\alpha_{1b}\theta_{1b} \geq \alpha_{sb}\theta_{sb} + \kappa\Omega(1)$; SSE with $\tau \in (0,1)$ exists iff $\alpha_{1b}\theta_{1b} = \alpha_{sb}\theta_{sb} + \kappa\Omega(\tau)$.

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 (9), $\tau = 0$ requires $\alpha_{0s}\theta_{0s}(V_1 - V_0 - V_s) \leq \kappa$. Hence $\tau = 0$ is consistent with SSE if $\alpha_{1b}\theta_{1b} \leq \alpha_{sb}\theta_{sb} + \kappa\Omega(0)$. 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 SSE; but they are not constants.*

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

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}, which is not too surprising since in some , but not all, search models (e.g., Lagos and Rocheteau 2009) certain results only depend on the product $\alpha\theta$. Still more generally, Proposition 2 gives a further adjustment for $\kappa > 0$. Crucially, for all these results, $\Omega(\tau)$ and the α's are endogenous.

The above is not meant to be a big advance over known results, but is a necessary step to what follows. To motivate the next step, note that despite (or maybe because of) the intuitive nature of RW it may be misleading. It does *not* say 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 μ_R is fundamentally inferior to μ_D, even if $\theta_{1b} = \theta_{sb}$ and $\kappa = 0$, and sometimes even if $\theta_{1b} > \theta_{sb}$ and $\kappa > 0$, middlemen can be active. Intuitively, M_1 may face an inferior meeting technology, suggesting τ should be 0; but if τ is small, their arrival rate can be big even with an inferior μ. Without claiming this is deep, we contend it leads to interesting insights below.

3 Meeting Technologies

The meeting process in RW is in some ways general but in other ways special: it allows flexible functional forms but restricts meetings between i and j to depend only on the mass of i and j, which is violated by many search models.[6] Indeed, it is violated in many middleman models that use *uniform random matching*, which means that the probability that i meets j is proportional to the fraction of type j in the market, which means, e.g., the rate at which M_1 meets B depends on all the N's, not just N_1 and N_b.

We now propose a market structure based on spatial separation that provides an interpretation of RW and uncovers some underlying assumptions. As shown in Figure 1, different types are located at distinct points represented by nodes on a triangle. There are three submarkets along the edges of the triangle: a direct market (DM) with B and S; a wholesale market (WM) with S and M_0; and a retail market (RM) with M_1 and B. For middlemen search is directed – those with inventory are only in RM, those without are only in WM. For B and S, everyone goes to the two closest markets, but not the third – it's just too far.[7]

[6] 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, which is inconsistent with the RW specification. The situation is similar in goods markets (e.g., Bethune et al. 2020).

[7] 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

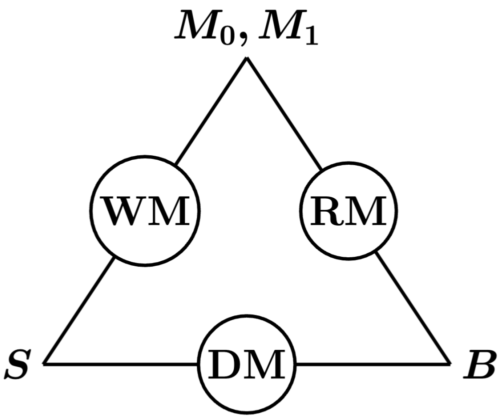

Figure 1: Market Structure

Assume B can participate in both the DM and the RM at the same time, and S can participate in both the DM and the WM 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 B households and two workers for S 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.[8] 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.

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 par-

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).

[8] 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.

ticular, DM meetings now depend on (N_b, N_s) and not other N's, WM meetings depend on (N_0, N_s) and not other N's, and RM meetings depend on (N_b, N_1) and not other N's consistent with RW. With a clearer understanding of this we can say much more about their model.

4 Beyond RW

Here we let $\kappa \to 0$ and suppose the meeting function in each submarket is a constant times the function $\mu(\cdot)$ (this is relaxed in the Appendix). Assumption 1 implies the constants must be the same in the WM and RM, but the constant in the DM 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* differences in the meeting technologies.

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

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

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 SSE.

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

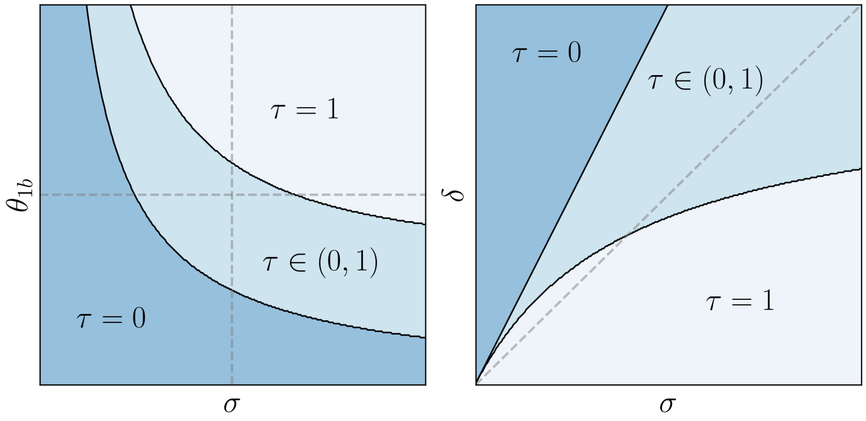

Figure 2: Equilibrium Regimes

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, and, by continuity, also when middlemen are somewhat worse. While Figure 2 is drawn for an example, the result is general.

Proposition 3 *Given $\delta\theta_{sb} > 0$, $\tau > 0$ is the SSE for $\sigma\theta_{1b} = \delta\theta_{sb}$ and for some $\sigma\theta_{1b} < \delta\theta_{sb}$. Moreover, $\tau = 1$ is the SSE 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 SSE 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 SSE if $\sigma\theta_{1b} > \delta\theta_{sb}\Phi(0)$. ■

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 $\tau > 0$. 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 more meetings leads to lower N, so an overall improvement in the meeting technology implies lower buyer-seller ratios (tightness) and, consequently, lower τ. That would not be apparent if one looked only at the RW result.

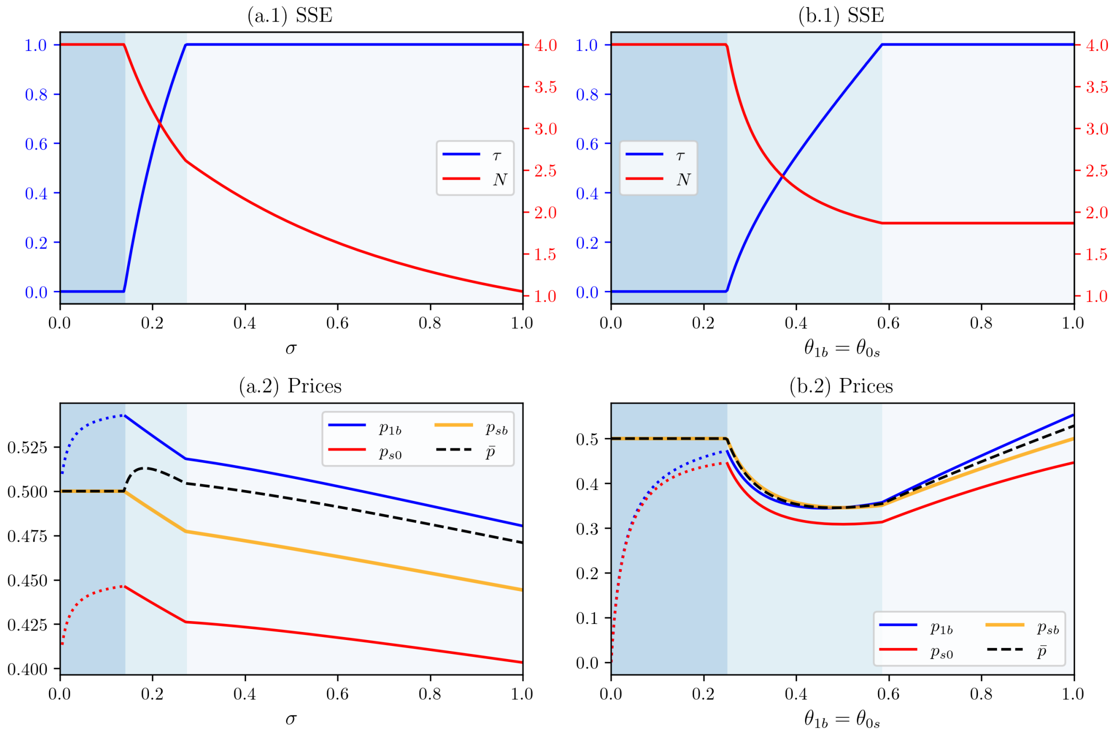

Figure 3: Parameter Effects

Figure 3 displays an example.[9] 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 prices fall, but when τ is big the impact of a marginal participant on competition diminishes and bargaining power dominates.

[9] This uses uniform random matching in each submarket, plus $E = 1$, $N_m = 5$, $\theta_{sb} = \delta = 0.5$, $u = 1$, $\kappa = 0$ and $r = 0.03$. There is nothing special about these parameters – all the examples we tried were fairly similar.

The average price paid by B, 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 RM relative to DM trade, and the RM price is above the DM price, $\bar{p}$ can increase. 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 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.[10]

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 $\theta_{1b} = \theta_{0s}$. That has been discussed elsewhere, and, in general, it is known that middlemen can increase or decrease welfare depending on details (e.g., Nosal et al. 2019), so we do not dwell on this.

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

[10] Having price dispersion non-monotone in frictions should be unsurprising to those who know search theory since Burdett and Judd (1983) get that. Having the average price non-monotone is harder. Lester (2011) gets it 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.

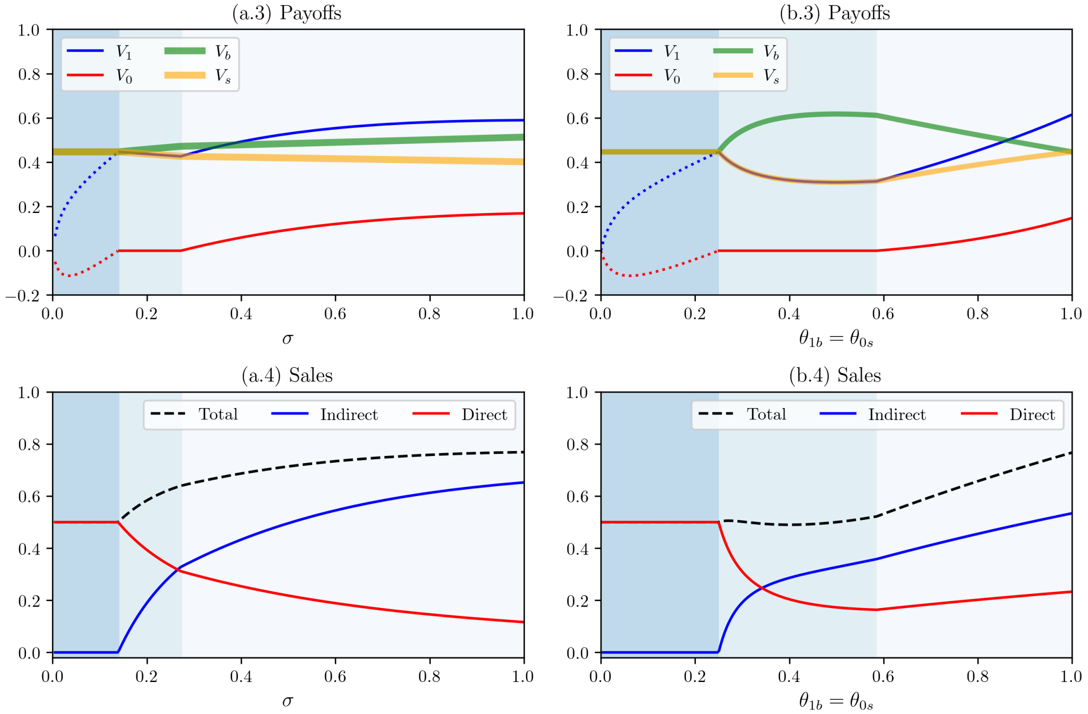

Figure 4: More Parameter Effects

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 as written.

Now consider imposing symmetry by letting S stay in the market forever, producing another good immediately after each sale. Then $\tau = 1$ is always consistent with SSE provided the middlemen meeting technology is not too inefficient and bargaining power is not too low. The reason is that now S has no opportunity cost to WM trade since upon giving M_0 a good S gets another one. So they trade as long as parameters for middlemen are not too unfavorable – e.g., θ_{1b} cannot be too low or M_1 cannot recover the cost of WM trade due to the usual holdup problem, given the payment to S is sunk when M_1 contacts B.[11] In any event,

[11] 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 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.

this version demonstrates plainly how middlemen can be fundamentally inferior to sellers yet still active in equilibrium.

5 Conclusion

We revisited a canonical search-and-bargaining model of middlemen. While RW made a big contribution by introducing this 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 be derived with heterogeneous bargaining powers or participation costs, but it is a characterization of an equilibrium outcome – middlemen activity – in terms of another equilibrium outcome – arrival rates. We characterized middlemen activity as well as arrival rates in terms of fundamentals, providing existence, uniqueness and comparative static results, some of which may be surprising (e.g., average prices need not fall when frictions fall), 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 as 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 clearly helped us better this model and 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 SSE. Otherwise, $\tau = 1$ if $g(\cdot) > 0$ or $\tau = 0$ if $g(\cdot) < 0$ is an SSE.

(Uniqueness) CRS of $\mu_D(\cdot)$ implies $\alpha_{bs} = \alpha_{sb} = \mu_D(1, 1)$ is constant in any SSE. CRS in $\mu_W(\cdot) = \mu_R(\cdot)$ (they are the same by Assumption 1) implies $\alpha_{1b,\tau=0} = \lim_{A \to 0} \mu_R\left(N_{\tau=0}/A, 1\right) > \mu_R\left(2N_{\tau=1}/N_m, 1\right) = \alpha_{1b,\tau=1}$. 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 WM when $\tau = 0$, and $\hat{\mu} \equiv \mu(1, 1)$ the arrival rate for B and S in DM. Since $\mu(\cdot)$ is strictly increasing, $\bar{\mu} > \hat{\mu}$. From (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," *IER* 43, 283-305.

M. Baye, J. Morgan and P. Scholten (2006) "Information, search, and price dispersion," *Handbook of Economics and Information Systems* 1, 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," *RES* 87, 691-720.

K.Burdett, and M. Coles (1997) "Marriage and Class," *QJE* 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," *JPE* 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," *JEP* 19, 139-58.

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

G. Gong and R. Wright (2024) "Middlemen in Search Equilibrium with Intensive and Extensive Margins," *IER* 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," *JPE* 97, 927-54.

R. Lagos and G. Rocheteau (2009) "Liquidity in Asset Markets with Search Frictions," *Econometrica* 77, 403-26.

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

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

18, 203-20.

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," *JET* 183, 876-906.

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

A. Rubinstein and A. Wolinsky (1987) “Middlemen,” *QJE* 102, 581-94.

M. Watanabe (2010) “A Model of Merchants,” *JET* 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” *JEL* 59, 90-148.