An Integrated Theory of Intermediation and Payments [*]

Marshall Urias[†]

Abstract

This paper develops an integrated theory of intermediation and payments in wholesale and retail goods markets. The model synthesizes the search-theoretic approach to intermedation with the New Monetarist approach to payments. I consider two margins of intermediation, inventory and entry, within pure credit and pure currency markets. In a pure credit economy, the equilibrium is generically inefficient due to an inventory holdup problem and search externalities. Improving the bargaining position of middlemen increases consumption and entry. In a pure currency economy, there is a two-sided holdup problem associated with middlemens' inventory choice and consumers' portfolio choice. This results in multiple steady state equilibria and a non-monotone response of consumption and entry to fundamentals. There exists a threshold nominal interest rate below which monetary policy is ineffective.

Keywords: monetary, intermediation, payments, search and matching

JEL Classification: E41, E42, E52

1 Introduction

There are few economic institutions as historically pervasive and essential as intermediated trade and payment arrangements. These two institutions are ubiquitous and naturally emerge to alleviate inherent frictions afflicting economies so that agents can realize gains from trade. Up until now, they have been studied independently even though the same set of frictions matter for the development of both. Rubinstein and Wolinsky [1987a] advocate "a basic model that describes explicitly the trade frictions that give rise to the function of middlemen," yet assume that these frictions are not relevant when agents must settle transactions. Monetary theorists venture to explain fiat money

[*] I thank Guillaume Rocheteau for his insights and suggestions which were essential. I thank Randy Wright, Alberto Trejos, David Andolfatto, and participants at the 2017 Annual Workshop of Money, Banking, Payments, and Finance at the Bank of Canada, 2017 Money, Banking and Asset Markets Conference at the Wisconsin School of Business, UCI Macroeconomic Seminar Series, and UCI Macroeconomic Reading Group for their thoughtful discussions.

[†] *Correspondance:* Peking University HSBC Business School, Nanshan, Shenzhen, China. Email: murias@phbs.pku.edu.cn

by acknowledging frictions in the transactions process, yet these frictions do not affect the ability of agents to trade directly with one another. In this paper, I develop a model which synthesizes the search-theoretic approach to intermediation with the New Monetarist economics of payments. The unified framework allows one to explore complementarities that exist between these institutions leading to several new insights.

The structure of exchange considered here is decentralized trade with pairwise meetings of agents where middlemen intermediate between consumers and producers. Middlemen have costly access to a search technology allowing them to procure inventory from producers in a wholesale market, and then sell inventory to consumers in a retail market. Prices and quantities are determined by sequential bilateral bargaining and depend on the form of payment arrangement used to settle transactions. The model exposits varying degrees of contract enforcement which generates different payment arrangements. First, I consider a pure credit economy where agents can fully commit to pay for current trades at a future date. Second, I consider the case where agents can strategically default on debt which endogenizes borrowing capacity. Third, I consider pure currency markets where credit is infeasible thus requiring agents to use a liquid asset to trade. The model addresses how the extent of intermediation, measured by entry and inventory, interacts with money and credit toward achieving desirable allocations, affects the response of the economy to changes in fundamentals, and influences the efficacy of monetary policy.

Equilibria are generically inefficient due to holdup problems and bargaining distortions. Even when there is perfect contract enforcement in both wholesale and retail markets, consumption falls short of the constrained efficient outcome due to an inventory holdup problem. Since middlemen must purchase inventory prior to meeting a consumer, this cost is sunk in the retail market leading to under-investment. Additionally, so long as middlemen do not possess total bargaining power in retail markets, they fail to realize the full rate of return on their inventory amplifying under-investment. Improving the bargaining position of middlemen—increasing its bargaining power or outside option—increases the number of active middlemen and improves the extensive margin of trade. The quantity per trade increases with retail bargaining power and decreases with wholesale bargaining power due to search externalities. Sequential bargaining in wholesale and retail markets allow the cost associated with the holdup problem to be divided between middlemen and producers. More specifically, the wholesale transaction internalizes the downstream search costs and distributes it between middlemen and producers. If a middleman has complete bargaining power in wholesale transactions, then the cost associated with the holdup problem is completely borne by the producer.

When credit is not feasible in retail transactions, there exists a two-sided holdup problem: one associated with inventory choices by middlemen and the other from portfolio choices by consumers. Consequently, there exist multiple steady state equilibria. Moreover, there is a non-monotone relationship between the quantity traded and the amount of entry. This is counter to a monetary economy without intermediation. In short, even when consumers choose to hold very large real balances, their consumption opportunities are still constrained by middlemen's inventory. When there are few middlemen, entry incentivizes consumers to hold more real balances and results in more trade. When there are many middlemen, search externalities cause middlemen to purchase fewer inventory which constrains consumption opportunities and results in less trade.

Due to this two-sided holdup, there are two regimes that dictate the response of the equilibrium to changes in fundamentals. Which regime the economy is in depends on the bargaining power of middlemen in the wholesale market and the cost of entry. Low wholesale bargaining power or high entry costs results in a regime characterized by few middlemen and hence consumer's liquidity constrains the allocation. High wholesale bargaining power or low entry costs results in a regime with many middlemen and hence inventory constrains the allocation. If the economy is in the former regime (liquidity constrained) then monetary policy behaves as usual: lower nominal rates increase consumers' real balances resulting in more consumption and entry. If, however, the economy is in the latter regime (inventory constrained), monetary policy is ineffective at some threshold nominal rate. Reducing the nominal interest rate does not affect real balances because consumers anticipate that they will not be able to use them given constrained inventory and monetary policy has no effect on the quantity traded or the entry of middlemen.

The rest of the paper is as follows. Sections **??** and **??** show that intermediation activities are relevant in real world economies and relate this paper to the literature on intermediation. Section **??** describes the intermediated economy and the role that money and credit play in facilitating trades. Section **??** establishes the welfare criterion against which decentralized equilibria is compared. Section **??** defines a decentralized equilibrium. Section **??** analyzes non-monetary equilibria, Section **??** non-monetary equilibria with strategic default and endogenous debt limits, and Section **??** monetary equilibria. Section **??** concludes and discusses how the framework can contribute to an intermediation theory of the firm.

2 Motivating Data

Apart from being anecdotally ubiquitous, intermediation activities constitute a non-trivial share of the U.S. economy. As a rough estimate, shares of GDP attributed to intermediation include retail trade (5.9 percent), wholesale trade (5.9 percent), finance and insurance (7.3 percent), transportation and warehousing (3 percent).[1] This estimate, a conservative one at that since it assumes zero value-added attributed to intermediation for all other industries, suggests intermediation activities account for approximately 22.1 percent of 2016 U.S. GDP. Moreover, disintermediation has not occurred. Figure 1 shows that this rough measure of intermediation has been relatively stable, just shy of one-quarter of the U.S. economy, for the last two decades.

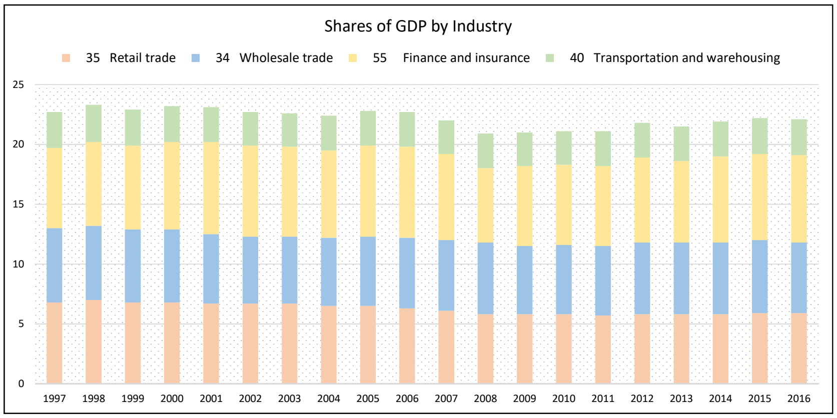

Figure 1: Intermediated Industries in U.S. GDP

Middlemen must exist because the benefits from intermediated trade are greater than those from direct trade. One explanation is that direct trade is very costly. Evidence by McGraw Hill (2013) estimates it takes on average 4.3 phone calls before a manufacturer finds a consumer at a total cost of just over $589. Although producers often pine for "cutting out the middleman," too often they forget that they still have to provide their function. If some producer bypasses intermediaries, it must incur the costs associated with distributing the good to end consumers. A survey of 200 local food producers in California by Brimlow (2016) found that 72 percent were selling out of

[1] The industry shares are taken from aggregate statistics published by the BEA.

the area to wholesale brokers. Producers surveyed claimed it was difficult to find local buyers or that marketing and advertising costs to sell locally would be too high. Middlemen provided a competitive advantage in delivering goods to final consumers. The benefits of intermediated trade may be amplified when goods are transported internationally. Intermediation has been a major driver of globalization bringing goods and services from local producers to international consumers.

The process by which middlemen match with producers and consumers is itself a productive process that uses resources. One of the most closely watched metrics to gauge retail performance is average inventory turnover; the reciprocal of which is days inventory outstanding (DOI). In a frictionless environment, middlemen could stock and sell inventories instantaneously and the DOI would be zero. However, it takes time to clear goods markets resulting in varying DOIs across industries and firms. There is a large degree of heterogeneity both across and within industries, but the average DIO in 2015 was 87.4 days. New logistic techniques, like just-in-time (JIT) inventory management, are developed to mitigate the costs associated with frictions. A model of intermediation in goods markets should take seriously the frictions that preclude the instantaneous purchase and resale of inventory.

Figure 2 plots two primary measures of U.S. retail establishments from 2008 to 2015. The extensive margin of intermediation is captures by the number of U.S. retail establishments while the intensive margin is captured by the ratio of inventories to sales. The graphs show that there exists moderate variation in intermediated trade over the business cycle and suggest a positive correlation between the two margins of trade.

Although technology has not led to disintermediation, as many tech guru prophets predicted, it has revolutionized retail payment technology. The means of executing quid pro trades using currency or digital substitutes is experiencing a rapid evolution and it is important to understand how the role of various retail payment instruments affect merchant behavior and vice versa. The 2015 Survey of Consumer Payment Choice (SCPC) finds that while there are nine identified payment instruments, consumers still predominantly use debit cards (32.5 percent of monthly payments), cash (27.1 percent), and credit cards (21.3 percent). The present model takes seriously the frictions that generate a need for middlemen and various payment arrangements.

Figure 3 plots the extent to which cash and credit are used in U.S. retail trades between 2008 and 2015. The graphs suggest that there was a preference toward using cash immediately following the

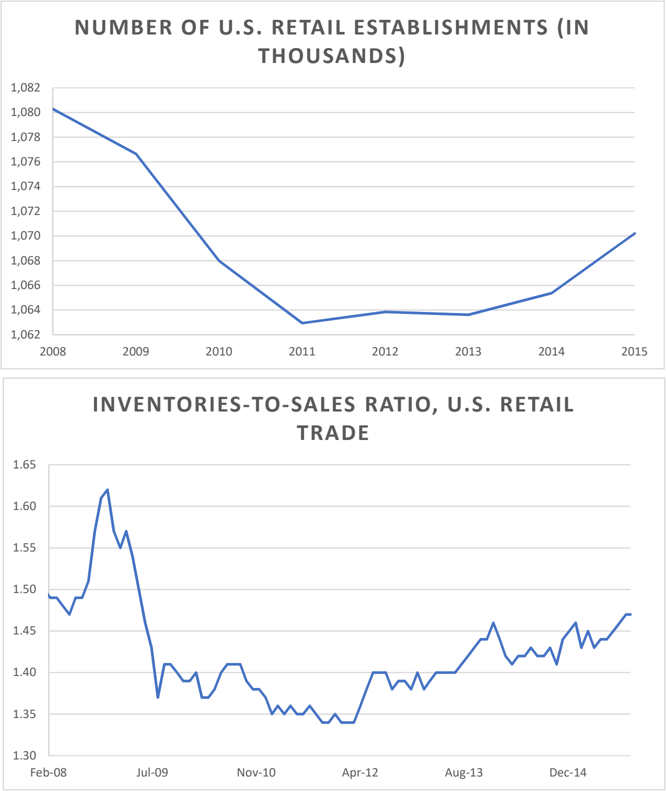

Figure 2: U.S. Retailers and Inventories

Great Recession and then credit gradually became more used in retail trades during the recovery.

This paper seeks to establish a connection between Figures 2 and 3. That is, is there a relationship between the payment instruments available to consumers and the extent of intermediated trade? To motivate this idea, consider the immediate aftermath of the Great Recession: with a collapse in the availability of credit, consumers were forced to use cash to execute retail trades. Scarce cash holdings limit the possible gains from trade thereby reducing retail sales, increasing inventories, and causing some firms to be unprofitable. This paper formalizes the relationship between cash, credit, and intermediated trades and includes as measure of credit availability as reflected in Figure 3 as well as endogenous entry of firms and investment in inventory reflected in Figure 2.

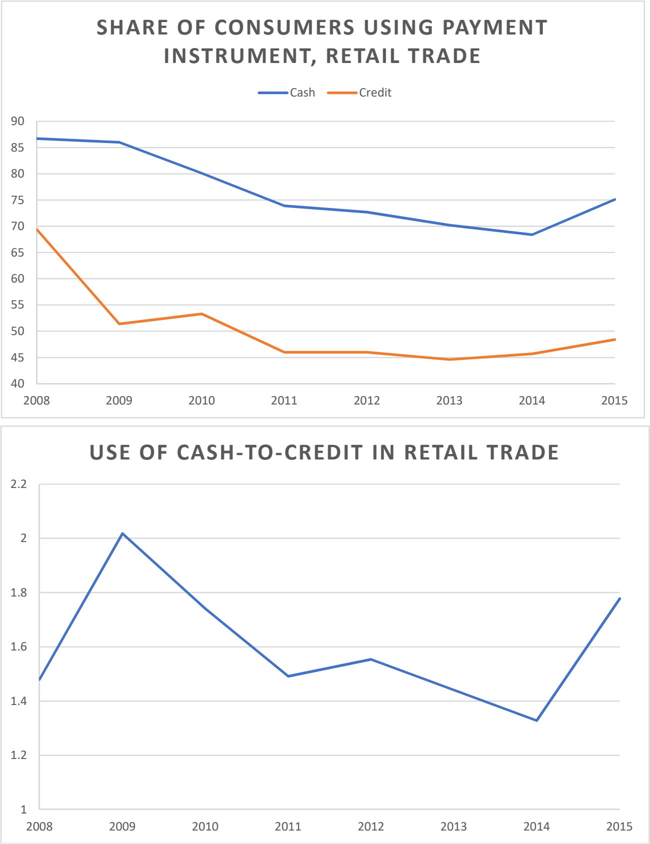

Figure 3: U.S. Payment Instruments in Retail Trade

3 Related Literature

The study of middlemen and how their presence influences market allocations dates to Rubenstein and Wolinsky (RW) who advocated “a basic model that describes explicitly the trade frictions that give rise to the function of middlemen.” RW modeled the exchange process and trading frictions between sellers, middlemen, and buyers thereby providing a framework for endogenizing the extent of intermediation and its effect on the distribution of gains from trade. Subsequent models, such as those by Nosal, Wong, and Wright (2011,2014,2016) (NWW), expand on RW to include production, search costs, Nash bargaining, and occupational choice. The present paper seeks to contribute the study of middlemen in the spirit of RW and NWW. The critical difference is that my paper integrates a rigorous theory of payment arrangements into a model of intermediation while retaining the enriching features found in NWW.

Also, different from NWW is that I consider an environment with infinite search and matching costs for direct trade, thereby generating an essential role for middlemen. This follows in the tradition of RW where middlemen are active in equilibrium only if they can match with consumers at a faster rate than producers. In RW, these matching rates are exogenous and so intermediated trade exists only if middlemen are endowed with a superior matching technology. In my model middlemen are indeed endowed with superior matching technology, but the rate at which intermediated trading opportunities arrive is endogenous and depends on the strategic choices of all agents. More specifically, a middleman's decision to enter the market depends on its relative bargaining position and likelihood of engaging in trade, which in turn is affected by the aggregate measure of active of middlemen.

There are various explanations for why middlemen are valuable. RW suggest that intermediation is a way of alleviating search frictions by providing more frequent consumption opportunities for consumers. Given some exogenous meeting process, some papers focus on the role of middlemen as guarantors of quality (Biglaiser 1993, Li 1998) while others suggest that middlemen can satisfy consumers' demand for varieties of goods whereas individual producers cannot [Johri and Leach, 2002] [Shevchenko, 2004]. Watanabe [2010] argues that middlemen have the advantage of inventory capacity relative to producers and that capacity constraints are an important determinant for the endogenous meeting rates. My paper takes the approach that middlemen alleviate search and matching frictions. This approach seems natural given that I want to jointly model payment arrangements and intermediation, both of which emerge from matching frictions and limited contract enforcement.

The literature on the coexistence of money and credit is robust. Lucas and Stokey (1987) propose a model where the distinction between trades executed with cash versus credit is exogenous. Subsequent work identified the fundamental frictions that generate a role for monetary exchange, e.g., lack of commitment and record-keeping (Kocherlokata 1998). Positing a costly record keeping technology endogenizes the composition of trades involving cash or credit (Camera and Li 2008, Bethune, Rocheteau, and Rupert 2015, Lotz and Zhang 2016). I follow the spirit of this literature in that I model monetary exchange in anonymous bilateral meetings and appeal to the inherent frictions in the environment to generate a role for liquid assets. However, the availability of credit to consumers remains exogenous.

4 Environment

Time is discrete and continues forever. Each period is divided into three stages where different transactions take place. The first stage agents trade is a wholesale market followed by a retail market in the second stage. The first two stages occur in decentralized markets (DM) where agents' trading opportunities arrive according to a random bilateral matching process. The arrival rate of a trading opportunity for an agent of type i with an agent of type j will be denoted α_{ij}. There exists a unique perishable and divisible retail good traded in the DMs denoted by q. During the third stage all agents meet in a centralized market (CM) where they produce, consume and exchange the numeraire good, denoted x, without trading frictions.

There are three types of agents: producers (P), consumers (C), and middlemen (M). Each type of agent is characterized by idiosyncratic preferences and technology. Producers have no desire to consume in the DM, but can produce the retail good. Consumers desire the retail good in the DM, but are unable to produce it. Middlemen do not desire the retail good, but can purchase it from producers in the first DM and resell it to consumers in the second DM. Additionally, any unsold inventory can be transformed at rate R into the numeraire during the CM.

In the first stage, a wholesale market opens where middlemen meet producers and purchase the retail good. Bilateral matches occur randomly according to arrival rates α_{pm}, α_{mp} such that a subset $\tilde{P} \subset P$ of producers are matched with $\tilde{M}_w \subset M$ middlemen. In the second stage, a retail market opens where consumers meet middlemen and purchase the retail good. Matches arrive according to α_{cm}, α_{mc} such that a subset $\tilde{C} \subset C$ of consumers are matched with $\tilde{M}_r \subset M$ middlemen. In the third stage, a centralized market opens where middlemen can transform unsold retail inventory into the numeraire at rate R and agents work to settle debts and adjust their liquidity holdings.

There are two payment systems available to consumers: money and credit. A fraction ω of trades are recorded and there exists perfect enforcement to guarantee repayment so that credit is a feasible payment system (Kocherlakota 1998). The remaining $1 - \omega$ fraction of matches are unmonitored precluding the use of credit so that only money can serve as payment.[2] I assume that there are no payment frictions in the wholesale market so that credit is always feasible between middlemen and producers. The terms of trade between a middleman and producer are denoted (q^w, b) where q^w is

[2] Although agents could use either money or credit in ω matches, they are payoff equivalent and so I restrict my attention to credit trades only.

the quantity a producer sells to a middleman and b is the debt issued by a middleman in exchange. The terms of trade between a consumer and middleman are denoted (q^r, p) where q^r is the quantity a middleman sells to a consumer and p is the corresponding payment, which is credit if the match is monitored with probability ω or money if the match in unmonitored with probability $1-\omega$.

Money is modeled as a perfectly divisible, intrinsically useless asset. Agents endogenously select to hold any non-negative amount of money allowing them to purchase the consumption good in the retail market. I assume that the quantity of money grows at a constant rate $M_{t+1} = \nu M_t$ and is injected by lump-sum transfers T to buyers. One unit of money m purchases ϕ units of the numeraire good in the centralized market. I call ϕ the value of money.

Agents maximize their discounted lifetime utility $\sum_{t=0}^{\infty} \beta^t U_t^j$, $j \in \{P, M, C\}$ where the period utility function of a producer, middleman, and consumer are given by,

$$\begin{aligned}
U^P &= -c(q_t) + x_t \\
U^M &= x_t \\
U^C &= u(q_t) + x_t
\end{aligned}$$

In the DM, consumers derive utility $u(q_t)$ from the retail good and producers incur cost $c(q_t)$ to produce it. In the CM, all agents enjoy linear utility in the numeriare good, where $x_t < 0$ is interpreted as the disutility of working to produce the numeraire. Middlemen only consume the numeraire and receive no utility from producing it for themselves since CM production costs are linear. To realize any positive utility, middlemen purchase the retail good from producers in the wholesale market and resell it to consumers in the retail market. Goods are non-storable between time periods and all agents discount future utility by a factor $\beta \in (0,1)$.

ASSUMPTION 1 *Utility $u(\cdot)$ and costs $c(\cdot)$ are C^2 functions defined on $\mathbb{R}_+$ and obey the usual properties: $u' > 0, u'' < 0, u(0) = 0, u'(0) = \infty, c' > 0, c'' > 0, c(0) = 0, c'(0) = 0$. Additionally, $\hat{q} < \tilde{q}$ where $u'(\hat{q}) = c'(\hat{q})$ and $u'(\tilde{q}) = R$.*

Differentiating types ex-ante makes it simple to introduce an extensive margin of trade. A subset of middlemen with measure n_t enter the wholesale market each period t at cost k. Normalizing the measure of buyers and sellers to one, bilateral matching guarantees that

$$\mu(n) = \alpha_{cm}(n) = n\alpha_{mc}(n)$$

$$\gamma(n) = \alpha_{pm}(n) = n\alpha_{mp}(n)$$

This specification allows search externalities in both wholesale and retail markets where trading opportunities depend on the ratio of middlemen to sellers and buyers n.[3]

ASSUMPTION 2 *Matching functions are homogeneous of degree one and exhibit standard properties:* $\mu'(n) > 0, \mu''(n) < 0, \mu(n) \leq \min(1, n), \mu(0) = 0, \mu'(0) = 1, \mu(\infty) = 1$ *and identical conditions on* $\gamma(\cdot)$*.*

5 Planner's Problem

I consider the problem of a social planner who each period chooses the measure n_t of active middlemen and an allocation $\{q_t^r(i), q_t^w(i)\}$ for all matched agents, $i \in \tilde{P} \cup \tilde{M}_w \cup \tilde{M}_r \cup \tilde{C}$, and $\{x_t(i)\}$ for all agents. The planner is constrained by the environment in the sense that he cannot choose the set of matched agents but only n_t, and then the sets are determined randomly in accordance with the matching technology. If the planner treats all agents identically, and confining attention to stationary allocations, the relevant period welfare function is given by

$$W_t = (2 + n)x + (\gamma(n)\mu(n)/n)u(q^r) - \gamma(n)c(q^w) - kn.$$

The first term is net consumption of the numeraire enjoyed by all agents. The second term is the utility of consumers (of measure 1) in the retail market who find a middleman holding inventory. The third term is the cost incurred by producers (of measure 1) who find a middleman. The fourth term is the cost of entry for middlemen (of measure n). The planner wishes to maximize

[3] Implicit in the description of the environment is that sellers are never matched directly with consumers. This can be interpreted as extreme matching frictions such that $\alpha_{cp} = \alpha_{pc} = 0$. In this sense, middlemen are essential to alleviate the extreme frictions placed on trade. Although stark, the purpose of this paper is to examine allocations in an environment with essential middlemen rather than derive the endogenous emergence of an intermediated sector.

$\sum_{t=0}^{t=\infty} \beta^t W_t$ subject to the following feasibility constraints,

$$(2+n)x \leq (\gamma(n)\mu(n)/n)R(q^w - q^r) + \gamma(n)(1 - \mu(n)/n)Rq^w$$
$$q^r \leq q^w$$

The first constraint states that net consumption of the numeraire can be no greater than unsold inventory transformed at rate R. The second constraint requires that an individual buyer can never purchase more than a middleman carries in inventory.

PROPOSITION 1 *The constrained efficient allocation* $(q^*, n^*) \in \mathbb{R}_+$ *solves the planner's problem and is given by,*

$$q^* = q^r = q^w \tag{1}$$
$$(\mu(n^*)/n^*)(u'(q^r) - R) = c'(q^w) - R \tag{2}$$
$$k = (\gamma(n^*)\mu(n^*)/n^*)'(u(q^r) - Rq^r) + \gamma'(n^*)(Rq^w - c(q^w)) \tag{3}$$

PROOF 1 *Maximizing* W_t *at each date, first order conditions for the planner's problem reveal that there are two potential solutions: one where the feasibility constraint* $q^r \leq q^w$ *binds and one where it does not. Assumption 1 rules out the non-binding case.*

Intuitively, (2) equates the marginal benefit of retail trade to the marginal cost of wholesale trade adjusted by the volume of meetings. The righthand side of (3) represents the value of retail and wholesale trades weighted by an entrants contribution to creating meetings; while the lefthand side is the cost of entering.

6 Decentralized Economy

Having described the constrained efficient allocation, I now consider a decentralized economy with intermediation and characterize stationary equilibria. I demonstrate that the efficient allocation never obtains, and the efficiency of the equilibria depends on the bargaining position of middlemen and the payment systems used.

6.1 Centralized Market

Consumers enter the CM with wealth comprised of debt and money $(b, m) \in \mathbb{R}^2_+$. Consumers choose how much to work in order finance consumption, repay debt, and adjust money holdings. They then enter the following period's DM which yields expected utility $V_t^C(m)$.

$$\max_{x,m'} W_t^C(b,m) = x + \beta V_{t+1}^C(m') \qquad s.t. \quad x + b + \phi_t m' = \phi_t m + T$$

Middlemen enter the CM with wealth comprised of net debt (credit from consumers less debt owed to producers), unsold inventory, and money $(b, q, m) \in \mathbb{R}^3_+$. They finance consumption of the numeraire using net wealth, transforming unsold inventory at rate R, and working. They then choose whether or not to enter the following period's DM with expected utility $V_1^M(m)$.

$$\max_{x,m} W_t^M(b,q,m) = x + \beta \max\{V_{1,t+1}^M, W_{t+1}^M\} \qquad s.t. \quad x + \phi_t m' = b + Rq + \phi_t m$$

A producer enters the CM with wealth comprised of credit and money Consumers enter the CM with wealth comprised of debt and money $(b, m) \in \mathbb{R}^2_+$ which it uses to finance its consumption of the numeraire.

$$\max_{x,m} W^P(b,m) = x + \beta V^P \qquad s.t. \quad x + \phi_t m' = b + \phi_t m$$

Substituting the budget constraints into their respective objective functions, the CM value functions for agents are given by the following:

$$W_t^C(m) = \phi_t m + T + \max_{m'} \left[-\phi_t m' + \beta V_{t+1}^C(m')\right] \tag{4}$$

$$W_t^M(q,m,b) = Rq + \phi_t m - b + \max_{m'} \left[-\phi_t m' + \beta \max\{V_{1,t+1}^M(m'), W_{t+1}^M(m')\}\right] \tag{5}$$

$$W_t^P(m,b) = b + \phi_t m + \max_{m'} \left[-\phi_t m' + V_{t+1}^P(m')\right] \tag{6}$$

Notice that all agents' CM value function are linear in wealth. When agents choose to acquire liquid assets, the portfolio decisions are history independent so that there is a degenerate distribution of asset holdings. This result is an artifact of quasi-linear preferences and delivers tractable results without sacrificing economic insight.

6.2 Retail Market

Having characterized the CM value functions, I move back one stage to the retail market where consumers and middlemen meet. A consumer entering the retail market finds a middleman with probability $\mu(n)$, and settles credit terms of trade (q^r, b^r) with probability ω or monetary terms of trade (q^r, d^r) with probability $1-\omega$. A consumer then enters the CM with its net wealth. Using the linearity of the CM value function (4), the expected utility to a consumer entering the retail market is given by

$$V_t^C = \mu(n)\left\{\omega[u(q^r) - b^r] + (1-\omega)[u(q^r) - \phi d^r]\right\} + W_t^C(m) \tag{7}$$

A middleman enters the retail market with some amount of inventory purchased from a producer, the corresponding debt, and money balances. With probability $\mu(n)/n$ he finds a consumer and with probability ω accepts credit and with probability $1-\omega$ only accepts cash. If the middleman does not find a consumer, he carries all unsold inventory and debt into the CM. Using CM value function (5) I have that,

$$V_{2,t}^M(q, b, \tilde{m}) = \frac{\mu(n)}{n}\left\{\omega[b^r - Rq^r] + (1-\omega)[\phi d^r - Rq^r]\right\} + W_t^M(q, b, \tilde{m}) \tag{8}$$

Equation (8) shows that a middleman's expected value in the retail market is the probability he finds a consumer times the value of the match, plus the guaranteed value of transforming unsold inventory in the CM, plus the continuation value of entering next period's wholesale market. Note that terms of trade in the retail market depend on what occurred in the wholesale market. If a middleman did not purchase any inventory in the wholesale market ($q = 0$) then he surely cannot sell anything in the retail market. More generally, the amount of inventory q constrains the set of feasible allocations in the retail market. This will be discussed more thoroughly when I define the bargaining sets.

6.3 Wholesale Market

I now move back one stage to the wholesale market where middlemen purchase inventory from producers. When a middleman enters the wholesale market he incurs entry cost k, meets a producer with probability $\gamma(n)/n$, executes terms of trade (q^w, b^w) and then enters the retail market. Otherwise he enters the retail market with zero inventory and zero debt.

$$V_1^M = (\gamma(n)/n)V_2^M(q^w, -b^w) + (1 - \gamma(n)/n)V_2^M(0, 0) - k$$

Using the retail value function (8) I have that,

$$V_{1,t}^M(\tilde{m}) = \frac{\gamma(n)}{n}\left\{\frac{\mu(n)}{n}\left(\omega[b^r - Rq^r] + (1-\omega)[\phi d^r - Rq^r]\right) + Rq^w - b^w\right\} + W_t(0, 0, \tilde{m}) - k \quad (9)$$

A middleman's expected value in the wholesale market is the expected value of acquiring inventory q^w with probability $\gamma(n)/n$. The value of holding inventory includes the guaranteed value of transforming it at rate R in the CM and repaying debts plus the value of carrying inventory into the retail market. Of course, the value of inventory in the retail market depends on the available payment instruments and resulting terms of trade. Suppose, for example, that the terms of trade are such that the consumer receives the entire value of surplus from a retail match. In this case, a middleman would only receive utility from transforming inventory in the CM and would never choose to operate in the retail market. This is an uninteresting equilibrium. To generate an equilibrium where consumers have an opportunity to consume and middlemen actually behave as intermediaries (buy and resell) it will be necessary to implement a bargaining protocol that gives the middleman some bargaining power in the retail market.

A producer entering the wholesale market finds a middleman with probability $\gamma(n)$, produces the good at cost $c(q)$, and receives credit to be settled in the CM.

$$V_t^P(\hat{m}) = \gamma(n)(-c(q^w) + b^w) + W_t^P(0, \hat{m}) \quad (10)$$

6.4 Bargaining Sets

I now characterize the set of allocations that are incentive feasible—the terms of trade which satisfy agents' participation constraints. The gains from retail trades crucially depend on the payment instrument used. For generality, I denote the payment made by a buyer by $p^r \in \{b^r, \phi d_r\}$ which may take the form of money or credit depending on the match.

First, I characterize the bargaining set that exists between a middleman and a consumer in the retail market. If an agreement is reached, a consumer's utility level is $u^C = u(q^r) + W^C(-p^r)$ and a middleman's utility level is $u^M = W^M(q - q^r, -b + p^r)$. If there is no agreement then a consumer receives utility $u_0^C = W^C(0)$ and a middleman receives utility $u_0^M = W^M(q, -b)$. Using the CM value functions, we can write the value of the surplus from a match as follows:

$$u^C - u_0^C = u(q^r) - p^r$$
$$u^M - u_0^M = p^r - Rq^r$$

A proposed trade is incentive feasible only if both agents earn non-negative surpluses from the agreement. The set of incentive feasible allocations is defined as $\Omega_r = \{(q^r, p^r) : Rq^r \leq p^r \leq u(q^r), q^r \leq q^w)\}$ where the payment instrument may be either money or credit $p^r \in \{b^r, \phi d^r\}$. The set can be constrained by two state variables: middlemen's inventory holdings q^w or consumer's real money balances ϕd^r which are predetermined when agents enter the match. A middleman can surely never sell more of the retail good than it has in inventory, and a consumer can never purchase more than the value of his real money holdings. Formally, the Pareto frontier of the bargaining set is described by

$$\max_{q^r, d^r} u^c = u(q^r) - p^r + u_0^C$$
$$s.t. \quad \phi d^r - Rq^r + u_0^M \geq u^M$$
$$s.t. \quad (q^r, d^r) \in [0, q^w] \times [0, m]$$

The jointly efficient outcome is $u'(\tilde{q}^r) - R = 0$ and $\tilde{p}^r = R\tilde{q}^r + u^M - u_0^M$. If a middleman carries too little inventory $q^w < \tilde{q}^r$ then a consumer will purchase all inventory, $q^r = q^w$, and compensate with payment $p^r = Rq^r + u^M - u_0^M$. If a consumer is liquidity constrained, $\phi m < \min\{Rq^r + u^M + u_0^M, R\tilde{q}^r + u^M + u_0^M\}$, then the consumer spends all money balances to acquire as much of the retail good as possible, $\phi m = Rq^r + u^M - u_0^M$.

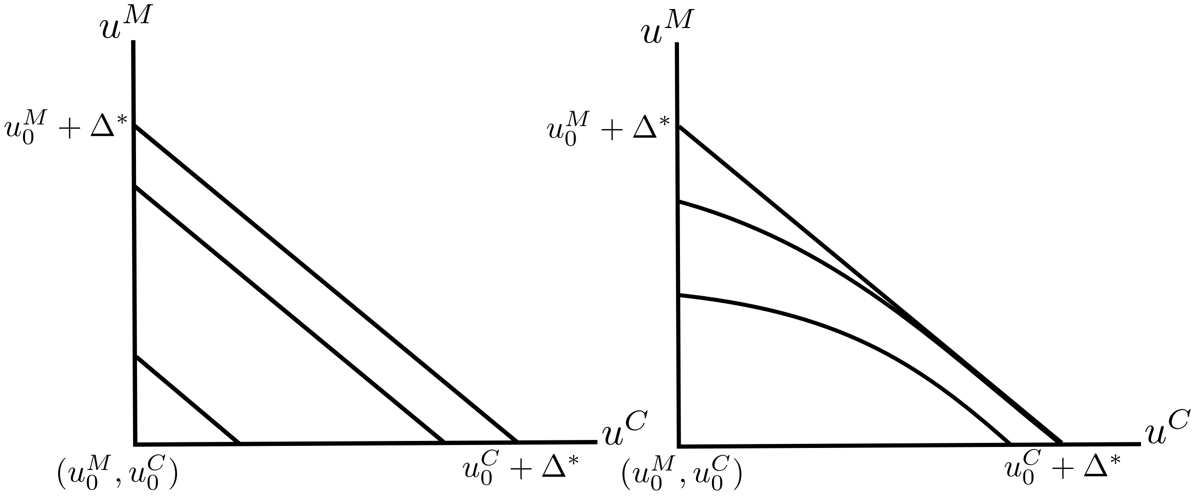

Figure 4: Pareto Frontiers for Ω_r

The maximum gains from trade depend on whether inventory or liquidity constrain the solution. If inventory is binding then the Pareto frontier is linear. If liquidity is binding, however, then the frontier is concave: $\frac{\partial^2 u^M}{\partial (u^C)^2} < 0$ if $\phi m - R\tilde{q} - (u^M - u_0^M) < 0$. Of course, if credit is available ($p^r = b^r$) then the liquidity constraint is irrelevant. Figure 4 depicts the two possible shapes of the frontier.

I now characterize the bargaining set that exists between a middleman and a producer in the wholesale market. If an agreement is reached, then a middleman receives utility $u^M = V_2^M(q^w, b^w)$ and the producer receives $u^P = -c(q^w) + W^P(b^w)$. If there is no agreement then the middleman gets $u_0^M = V_2^M(0,0)$ and the producer gets $u_0^P = W^P(0)$. Using the CM value functions, we can write the value of the surplus from a match as follows:

$$u^M - u_0^M = \pi(q^w) - b^w$$
$$u^P - u_0^P = -c(q^w) + b^w$$

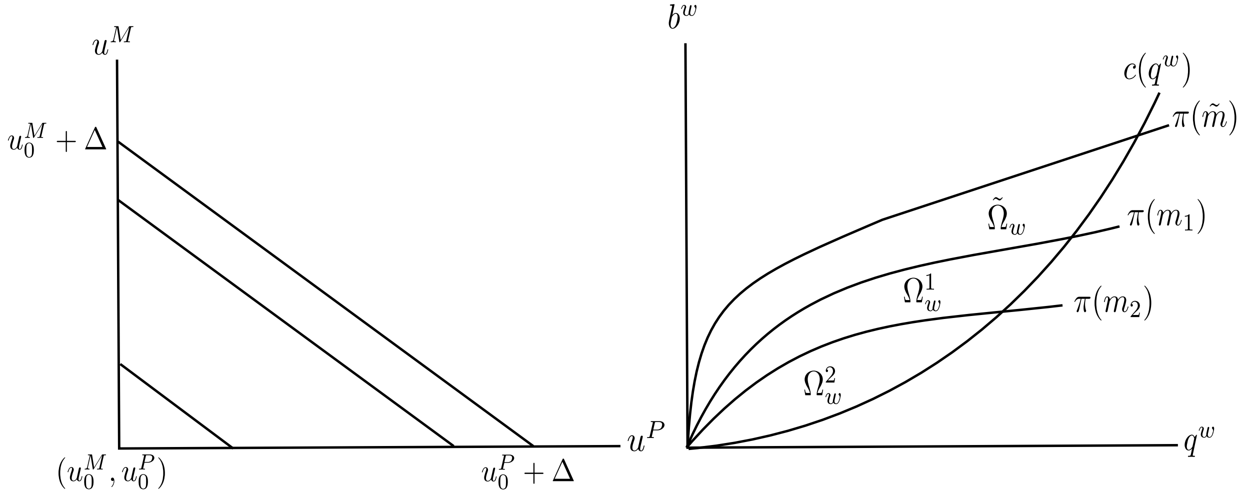

Figure 5: Incentive Feasible Set and Pareto Frontier for Ω_w

where $\pi(q^w) = \frac{\mu(n)}{n}\left[\omega(-Rq^r(q^w) + b^r(q^w)) + (1-\omega)(\phi d^r - Rq^r)\right] + Rq^w$ is the expected surplus from retail trade. The set of incentive feasible allocations is thus defined as $\Omega_w = \{(q^w, b^w) : c(q^w) \leq b^w \leq \pi(q^w)\}$. Note that once the amount of inventory exceeds the jointly efficient retail quantity, the marginal benefit of inventory is simply its transformation value in the CM. That is, the upper contour of the bargaining set is linear with slope R for all $q > \tilde{q}$. Consumers' portfolio choice will affect the size of the total surplus in the retail market, and the terms of trade will dictate how that surplus is divided. Therefore, the feasible set Ω_w depends on both consumers' real money holdings and the bargaining protocol. [4] Also note that greater entry induces a congestion effect in the retail market shrinking the set of incentive feasible trades. As $n \to \infty$ the feasible trades must satisfy $c(q^w) \leq b^w \leq Rq^w$ indicating that middlemen only realize value from transforming inventory into the numeraire. Alternatively, as $-Rq^r(q^w) + b^r(q^w) \to 0$ we again have that $c(q^w) \leq b^w \leq Rq^w$. Whether the efficient quantity traded is incentive feasible $q^* \in \Omega_w$ depends on the share of retail trade surplus a middleman receives. Of course, if $c(q^*) \leq Rq^*$ then the efficient quantity is incentive feasible even when the middleman's share of retail surplus approaches zero; although this is not true in general and largely depends on the rate R at which a middleman can transform unsold inventory.

[4] Suppose, for example, that middlemen receive zero share of retail surplus. Then the jointly efficient outcome of wholesale trades would reduce to $R = c'(q^w)$; but this corresponds to middlemen choosing not to enter the retail market and simply transform inventory into the numeraire. Any equilibrium with active middlemen in the retail market requires that middlemen receive a non-zero share of retail surplus.

The Pareto frontier of the bargaining set is defined by the following:

$$\max_{q^w,b^w} u^M = \pi(q^w) - b^w + u_0^M$$
$$s.t. \quad -c(q^w) + b^w + u_0^P \geq u^P$$

The jointly efficient outcome is given by the solution to the following,

$$(\mu(n)/n)\frac{\partial(-Rq^r(q^w) + b^r(q^w))}{\partial q^w} + R = c'(q^w)$$
$$b^w = u^P - u_0^P + c(q^w)$$

The jointly efficient allocation equates the marginal benefit to a middleman of acquiring inventory to the cost of producing that inventory. The marginal benefit to a middleman is the surplus received in the retail market with probability $\mu(n)/n$ plus the ability to transform any unsold inventory at rate R with probability one. With the optimal quantity determined, debt is issued by the middleman to compensate the producer for its cost of production and provide some surplus.

The Pareto frontier is linear and strictly decreasing in the share of surplus received by middlemen in the retail market,

$$u^M - u_0^M = (\mu(n)/n)[-Rq^r + p^r] + Rq^w - c(q^w) - (u^P - u_0^P)$$

6.5 Free Entry of Middlemen

A middleman participates in the retail market if $V_2^M(q^w, -b^w) \geq W^M(q^w, -b^w)$. This is equivalent to

$$\frac{\mu(n)}{n}\left\{\omega[b^r - Rq^r] + (1-\omega)[\phi d^r - Rq^r]\right\} > 0$$

which says that a middleman must receive a non-negative expected surplus from participating in the retail market.

A middleman chooses to search in the wholesale market if the value of doing so is at least a great as the cost of entry, $V_1^M \geq 0$. Using the value functions this is equivalent to

$$\frac{\gamma(n)}{n}\left\{\frac{\mu(n)}{n}\{\omega[b^r - Rq^r] + (1-\omega)[\phi d^r - Rq^r]\} + Rq^w - b^w\right\} \geq k \tag{11}$$

The value of searching in the wholesale market is the value of selling inventory in the retail market with probability $(\mu(n)/n)$ plus the value of transforming inventory in the centralized market with probability one. Notice that it may be profitable for a middleman to acquire inventory in the wholesale market even if it expects no surplus from the retail market. This occurs when $(\gamma(n)/n)(Rq^w - b^w) \geq k$ – inventory holding costs must be sufficiently low to induce entry in the wholesale market if there is no surplus from the retail market. Any surplus from retail trades relaxes the condition for entry.

7 Non-Monetary Equilibrium

An equilibrium is defined as $\{q^w, q^r, b^w, b^r, d^r, n\}$ and money holdings $\{m_j\}, j = C, M, P$ where a bargaining protocol determines the terms of trade, free entry determines the measure of active middlemen, and portfolio choices determine money holdings. As was mentioned previously, the availability of credit versus money impacted the incentive feasible trades in both retail and wholesale markets and affects the entry decision of middlemen. For expositional clarity, I consider the limiting cases of a pure credit economy ($\omega = 1$) and a pure currency economy ($\omega = 0$). It is straightforward to characterize all remaining equilibria as the convex combination of these two limiting cases.

First, I consider non-monetary stationary equilibria ($\omega = 1$). I begin with the retail market where a consumer meets a middleman. The terms of trade will be determined by Kalai proportional bargaining,

$$\begin{aligned}
&\max_{q^r, b^r} u(q^r) - b^r \\
&s.t. \quad u(q^r) - b^r = \frac{\theta_r}{1-\theta_r}(-Rq^r + b^r) \\
&s.t. \quad q^r \leq q^w
\end{aligned}$$

where θ_r denotes the bargaining power of a consumer. The unconstrained solution obtains where the marginal benefit of consumption equals the marginal opportunity cost of the sale $u'(\tilde{q}) = R$ and the corresponding transfer is $\tilde{b} = \theta_r R\tilde{q} + (1-\theta_r)u(\tilde{q})$. If a middleman holds too little inventory $q^w < \tilde{q}$ however, then the solution is constrained such that a consumer purchases all inventory and issues the corresponding amount of debt,

$$q^r = q^w \tag{12}$$

$$b^r = \theta_r Rq^r + (1-\theta_r)u(q^r) \tag{13}$$

Under proportional bargaining, the surplus received by either agent monotonically increases as the bargaining set expands. An extra unit of inventory held by a middleman (relaxing the constraint) generates extra surplus up to the jointly efficient allocation. That is, $\partial S_r/\partial q^w = u'(q^w) - R$ if $q^w < \tilde{q}$ and is zero otherwise. The terms of trade in the wholesale market, where a middleman purchases inventory from a producer, are settled according to,

$$\max_{q^w, b^w} (\mu(n)/n)[-Rq^r(q^w) + b^r(q^w)] + Rq^w - b^w$$
$$s.t. \quad (\mu(n)/n)[-Rq^r(q^w) + b^r(q^w)] + Rq^w - b^w = \frac{\theta_w}{1-\theta_w}(-c(q^w) + b^w)$$

where θ_w denotes the bargaining power of a middleman. The solution to the above program is

$$(\mu(n)/n)(1-\theta_r)(u'(q^w) - R)^+ + R = c'(q^w) \tag{14}$$

$$b^w = (1-\theta_w)\left((\mu(n)/n)(1-\theta_r)(u(q^w) - Rq^w) + Rq^w\right) + \theta_w c(q^w) \tag{15}$$

where $x^+ = \max\{0, x\}$. The marginal benefit from an extra unit of inventory is the expected surplus it brings to a retail match plus its guaranteed recycle value. The optimal amount of inventory equates this marginal benefit to the marginal cost borne by a producer.

PROPOSITION 2 *For a given level of entry n, the amount of inventory purchased in the decentralized equilibrium with credit is less than the jointly efficient quantity in retail trades, $q^w < \tilde{q}$ and less than the first-best allocation, $q^w < q^*$, for all $\theta_r > 0$.*

The first result is due to a hold-up problem: since a middleman is required to purchase inventory prior to meeting a consumer, this cost is sunk during bargaining in the retail market. Consequently,

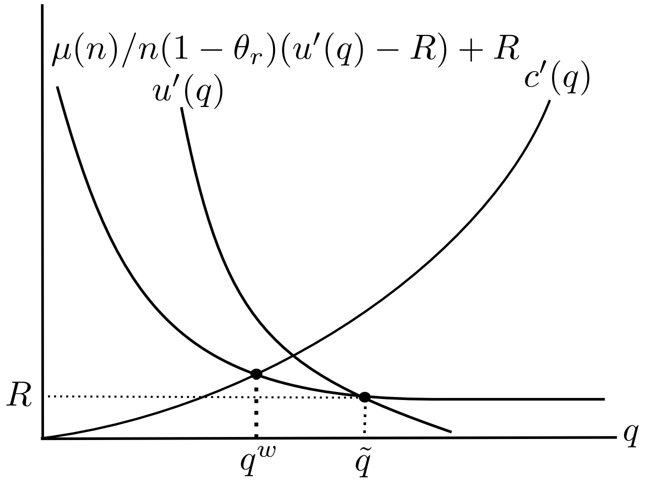

Figure 6: Inventory Purchases in Pure Credit Economy

a middleman will never purchase enough inventory to satiate consumer demand. The second result follows because a middleman does not receive the full value of his investment in inventory in the subsequent retail market so long as $\theta_r > 0$. A graphical description is provided in Figure 6.

Middlemen endogenously choose to participate in the wholesale market at cost k. Free entry implies $V_1^M = W^M(0) = 0$. Using (7) I have that,

$$\frac{\gamma(n)}{n}\theta_w\left(\frac{\mu(n)}{n}(1-\theta_r)(u(q^w)-Rq^w)+Rq^w-c(q^w)\right)=k \tag{16}$$

Given the quantity traded in the wholesale market, the measure of middlemen n will adjust such that the value of entering the wholesale market and the value of not entering are equated to zero. Notice that a necessary assumption to guarantee $n > 0$ is that $\theta_w\left((1-\theta_r)(u(q^w)-Rq^w)+Rq^w-c(q^w)\right) \geq k$ where q^w is determined by (14). This requires $\theta_w > 0$ and the constraint is most slack when $\theta_r = 0$.

An equilibrium jointly describes the terms of trade and measure of operative middlemen using (12)-(16). Notice that there exist complementarities between the extensive and intensive margins. The probability of a trading opportunity depends on the measure of active middlemen which in turn affects the quantity of inventory purchased in the wholesale market. Observing (14), greater entry induces lower expected utility in the retail market which exacerbates under-investment and hence reduces retail consumption.

An equilibrium can be summarized by equations (14) and (16) which jointly determine (q^w, n).

By (14) there is an inverse relationship between the measure of active middlemen and the quantity of retail consumption. Higher entry causes a congestion effect which lowers the expected surplus in the retail market and leads to under-investment of inventory. Define $\bar{q}$ such that $(1-\theta_r)(u'(q)-R) = c'(q)-R$ which corresponds to the case where $n \to 0$. As the probability of finding a trading partner in the retail market approaches one the quantity traded increases up to $\bar{q}$ which is still less than the first best because middlemen do not realize the full return of their investment in inventory as long as $\theta_r > 0$. Define $\underline{q}$ such that $c'(q) = R$ which corresponds to the case where $n \to \infty$. As entry becomes unbounded the probability of finding a match approaches zero so that a middleman only realizes return on inventory by transforming it into the numeraire. Free entry condition (16) shows a positive relationship between entry and the quantity of the consumption good traded. To induce entry, a middleman must be compensated for the lower probability of a match with a greater quantity traded per match. Equilibrium (q^w, n) occurs at the intersection of (14) and (16). The quantity of debt issued is then determined by (13) and (15) where it is clear from (12) that $q^r = q^w$.

PROPOSITION 3 *The decentralized equilibrium can never reach the socially efficient allocation.*

PROOF 2 *The first best quantity traded obtains only where $\theta_r = 0$ so that middlemen realize the full return on their inventory. However, optimal entry occurs for a version of the Hosios condition where $\theta_r = -(\gamma(n)/\gamma'(n))((\mu(n)/n)'/(\mu(n)/n))$ and $\theta_w = n\gamma'(n)/\gamma(n)$ obtained from comparing* (16) *to* (3)*. These two optimality conditions are contradictory therefore the socially efficient (q, n) cannot be reached.*

PROPOSITION 4 *Decreasing the bargaining power of middlemen in the retail market results in less entry and less trade, $\partial n/\partial\theta_r < 0, \partial q/\partial\theta_r < 0$. Increasing the bargaining power of middlemen in the wholesale market results in more entry but less trade, $\partial n/\partial\theta_w > 0, \partial q/\partial\theta_w < 0$. Increasing the cost of entry results in less entry but more trade, $\partial n/\partial k < 0, \partial q/\partial k > 0$. A decrease in inventory holding costs increases entry and has an ambiguous effect on quantity traded, $\partial n/\partial R > 0$.*

If the extensive margin were shut down (exogenous n) then altering the bargaining power in the wholesale market would have no effect on the quantity traded; it would simply adjust the division of surplus between producers and middlemen. However, when entry is endogenous middlemen

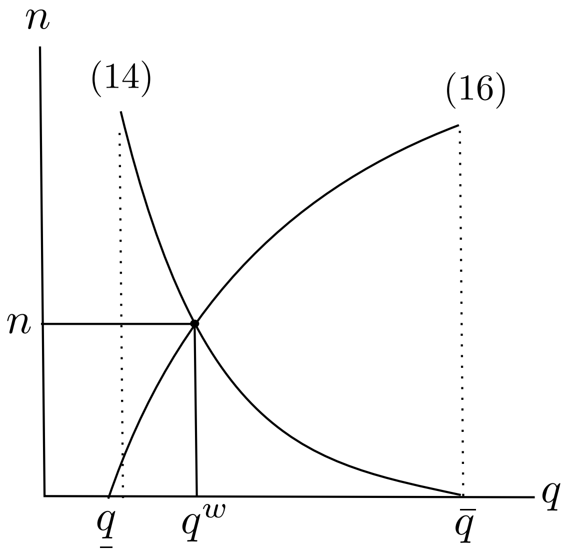

Figure 7: Equilibrium in (q, n)

internalize the higher share of expected surplus resulting in more entrants. More entrants decrease the probability of a retail match which incentivizes middlemen to purchase less inventory and thus engage in less trade. Also interesting is that higher entry costs can increase the quantity traded. Entry costs have no effect on the division of surplus as can be seen from (14); however, they do affect ex-ante profits seen from (16). To compensate for the lower probability of a retail match due to fewer middlemen, a greater quantity must be traded in equilibrium. The effect on entry from a decrease in on inventory holding costs is ambiguous. Lower inventory costs improve a middleman's outside option resulting in more trade in retail matches; (14) shifts northeast. Concurrently, an improved outside option encourages more entry which congests the wholesale market and decreases inventory purchases; (16) shifts northwest.

PROPOSITION 5 *The intermediation spread for middlemen is strictly positive and given by,*

$$(b^r - b^w)(q) = (u(q) - Rq)(1 - \theta_r - (1-\theta_r)(1-\theta_w)\mu(n)/n) + \theta_w(Rq - c(q)).$$

The spread is increasing in the bargaining power of middlemen $\partial(b^r - b^w)/\partial\theta_w > 0, \partial(b^r - b^w)/\partial\theta_r < 0$*, the amount traded* $\partial(b^r - b^w)/\partial q > 0$*, inventory holding costs* $\partial(b^r - b^w)/\partial R > 0$ *and the measure of middlemen* $\partial(b^r - b^w)/\partial n > 0$*.*

Of interest is the proportion of retail surplus captured by a middleman: $(1 - \theta_r - (1-\theta_r)(1-$

$\theta_w)\mu(n)/n)$. The first term captures the primitive bargaining power of middlemen in retail trade, whereas the second term reveals the interaction between wholesale and retail trade. Suppose, for example, that a middleman has all bargaining power in wholesale trades $\theta_w = 1$. In this case, the producer is forced to internalize not only its own production costs, but also the search costs associated with the retail market. That is, the wholesale transaction internalizes the downstream search costs and distributes it between middlemen and producers. This mechanism underlies the intuition for why $\partial(b^r - b^w)/\partial n > 0$. More entry decreases the expected value of a retail match, and therefore requires a smaller payment in wholesale trades. Concurrently, more entry does not affect the terms of trade in retail matches. A middleman can extract up to the full surplus $u(q) - c(q)$ when $\theta_r = 0, \theta_w = 1$.

8 Limited Commitment

Thus far I have assumed that credit is perfect. That is, there exists a record keeping technology and enforcement mechanism that replicates perfect memory and ensures debt repayment. Now suppose that such an enforcement technology does not exist, and so repayment of debt must be self-enforcing. Buyers will be allowed the possibility of strategic default, but understand that their actions are publicly recorded and punishment for default is exclusion from all future credit trades.

I begin with the retail market, and denote $\bar{b}^r$ the consumer's debt limit which is the maximum amount that a buyer is willing to repay. The consumer will have an incentive to repay his debt in the CM if and only if $-b^r + \beta V^C \geq 0$. The sum of the buyer's current and continuation payoffs if he repays his debt must be greater than the continuation (autarkic) payoff of zero if he defaults. The debt limit is thus defined as,

$$\bar{b}^r = \frac{\mu(n)}{\mu(n) + r} u(q^r) \tag{17}$$

and the set of incentive feasible allocations in the retail market is given by $\Omega_r^{lc} = \{(q^r, b^r) : Rq^r(q^w) \leq b^r \leq \bar{b}^r\}$.[5] Compared to full commitment, the set of feasible trades is strictly smaller. Moreover, the debt limit is increasing with the measure of active middlemen. More middlemen increase the frequency of trading opportunities which makes having access to credit more valuable.

[5] There exist a continuum of stationary credit equilibria indexed by debt limits $\bar{b} < \bar{b}^r$ supported by self-fulfilling beliefs. I restrict my attention to the "not-too-tight" borrowing constraint which is sufficiently tight to prevent default but not too tight so as to leave unexploited gains from trade.

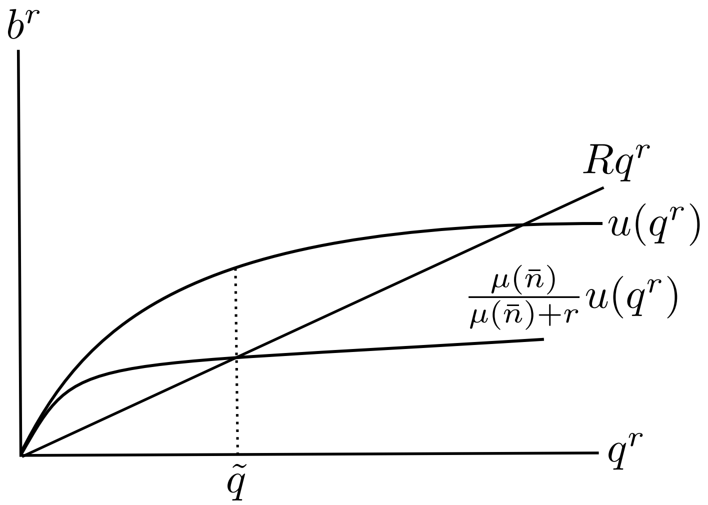

Figure 8: Limited Commitment Retail Bargaining Set

If the jointly efficient quantity lies in the incentive feasible set, $\tilde{q} \in \Omega_r^{lc}$, depends on (q^w, n, β). That is, for any given discount factor there exists a threshold level of entry such that if there are too few middlemen then the jointly efficient quantity is not incentive feasible. The intuition is that if there are too few middlemen in the retail market, then exclusion from future retail trades is not punishing enough to induce debt repayment.

I now consider the wholesale market, and denote $\bar{b}^w$ the middleman's debt limit defined by $-\bar{b}^w + \beta V_1^M = 0$, or written explicitly,

$$\bar{b}^w = \frac{(\gamma(n)/n)(\mu(n)/n)(-Rq^r + b^r) + (\gamma(n)/n)Rq^w - k}{r + \gamma(n)/n} \tag{18}$$

and the set of incentive feasible allocations in the wholesale market is given by $\Omega_w^{lc} = \{\bar{b}^w \leq b^w \leq c(q^w)\}$. If the efficient amount of inventory is incentive compatible depends on (n, β, k). The debt limit is decreasing in the measure of operative middlemen n, increasing in their patience β, and decreasing in the cost of entry k. For a given (β, k) there exists a threshold level $\bar{n}_w$ such that for all $n \geq \bar{n}_w$ there is no b^w that can support the efficient quantity trade. Intuitively, too many middlemen congest the market which reduces the benefit of avoiding autarky.

I continue to assume that terms of trade are settled by proportional bargaining. In a retail match, the jointly efficient quantity $\tilde{q}$ is purchased if $\bar{b}^r > \theta_r R\tilde{q} + (1-\theta_r)u(\tilde{q})$. Otherwise, the consumer borrows up to the debt limit and purchases the maximum amount that a middleman is willing to sell, $\bar{b}^r = \theta_r R q^r + (1-\theta_r)u(q^r)$. In a wholesale match, a middleman will purchase the jointly efficiency quantity given by (14) if $\bar{b}^w > b^w$ where b^w is given by (15). Otherwise, the middleman borrows up to the debt limit and purchases as much as a producer is willing to sell in exchange for $\bar{b}^w$.

In the following sections, I further relax the notion that agents can commit and introduce a role for a medium of exchange.

9 Monetary Equilibria

In this section, I investigate the role that money plays in facilitating trade within an intermediary sector. I assume that money is necessary in retail market transactions due to anonymity and lack of record keeping, and that credit is feasible in the wholesale market for simplicity.[6] Money is modeled as a perfectly divisible, intrinsically useless asset. Agents endogenously select to hold any non-negative amount of money allowing them to purchase the consumption good in the retail market. I assume that the quantity of money grows at a constant rate $M_{t+1} = \nu M_t$ and is injected by lump-sum transfers T to buyers. One unit of money m purchases ϕ units of the numeraire good in the centralized market. I call ϕ the value of money.

The critical difference is the terms of trade in the retail market. Since credit is not feasible between middlemen and consumers, the terms of trade in the retail market (q^r, d^r) indicate a quantity of good exchanged for some amount of fiat money d^r. In the CM all agents exchange money and goods. In principle, any type of agent can choose to accumulate money in the CM. As we will see, however, only consumers realize liquidity value from holding money in the retail market.

For comparability with the pure credit economy, I continue to settle the terms of trade according to proportional bargaining. In the retail market we have,

$$\max_{q^r,d^r} u(q^r) - \phi d^r \quad s.t. \quad u(q^r) - \phi d^r = \frac{\theta_r}{1-\theta_r}(-Rq^r + \phi d^r)$$

$$s.t. \quad q^r \leq q^w, \quad d^r \leq m$$

[6] We may imagine that producers are sophisticated in the sense that they are able to record and recognize members of the intermediary sector. That is, each producer has technology which assigns a name to each middleman and can find said middleman in the CM to collect on debts.

As before, the unconstrained solution is such that

$$u'(\tilde{q}) = R$$
$$\phi\tilde{d} = \theta_r R\tilde{q} + (1-\theta_r)u(\tilde{q})$$

Now there are two constrained solutions. If inventory is insufficient we have that,

$$q^r = q^w$$
$$\phi d^r = (1-\theta_r)u(q^r) + \theta_r R q^r$$

If money holdings are insufficient we have that,

$$\phi m = (1-\theta_r)u(q^r) + \theta_r R q^r \tag{19}$$

There are two reasons why the jointly efficient trade may not obtain. First, a middleman may purchase too little inventory since this investment decision is made ex-ante and bargaining occurs ex-post. Second, a consumer may hold too few real money balances—also the consequence of an ex-ante portfolio decision. In the former case, a consumer purchases all available inventory in exchange for real money balances that gives the middleman a fraction $(1-\theta_r)$ of the joint surplus. In the latter case, a consumer spends all real balances to purchase inventory that gives the consumer a fraction θ_r of the surplus.

A consumer's choice of money holdings is given by (4) where I substitute out V^C using (7),

$$\max_m -(\phi_t - \beta\phi_{t+1})m + \beta\mu(n)[u(q^r(q^w, m)) - \phi d^r(q^w, m)] \tag{20}$$

Notice that if $\phi_t/\phi_{t+1} < \beta$ then there is no solution to (1) since consumers would demand infinite money balances. If $\phi_t/\phi_{t+1} = \beta$ then the cost of holding money is equated to the rate of time preference and agents' choice of money holdings is enough to purchase $\tilde{q}$ and is not unique. Finally, if $\phi_t/\phi_{t+1} > \beta$ then money is costly to hold and buyers do not carry more money balances than they expect to spend in the retail market and the solution is unique.

I restrict my attention to stationary equilibria (i.e. $q_t = q_{t+1} = q$) which requires that aggregate

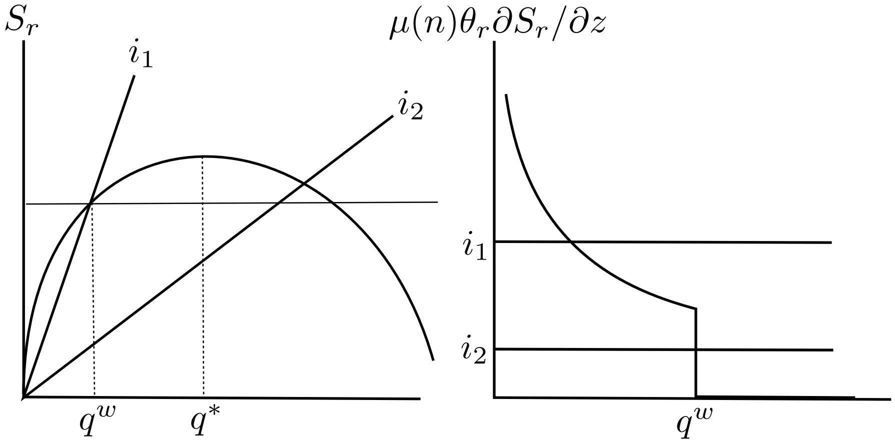

Figure 9: Consumer's Portfolio Decision

real money balances are constant over time: $\phi_t M_t = \phi_{t+1} M_{t+1}$. There is thus a one-for-one mapping between the rate of money growth and the rate of inflation: $\phi_{t+1}/\phi_t = 1/\nu$. Considering stationary equilibria, and using the proportional bargaining outcome, consumers' choice of real money balances follows

$$\max_{q^r} -iz(q^r) + \mu(n)\theta_r S_r(q^r) \tag{21}$$

where $(1+i) = (1+r)\nu$ is the nominal interest rate on an illiquid bond, $z = \phi m$, $z(q) = (1-\theta_r)u(q)+\theta_r Rq$ is the mapping from real balances to consumption given by the proportional bargaining solution, and $S_r(q) = u(q) - Rq$ is the total surplus from retail trade. A consumer weighs the cost of holding money $-iz$ against the liquidity value it brings in the retail market $\mu(n)\theta_r S_r(q^r(z))$. T o guarantee that the problem remains concave and admits a unique maximum it must be the case that $\theta_r/(1-\theta_r) > i/\mu(n)$. For a given level of entry, the buyer must have enough bargaining power in the retail market for money to be valued in equilibrium. The consumer's problem is represented graphically in Figure 9. Consumers realize positive marginal benefit from holding additional money balances up to the threshold q^w representing a middleman's inventory. If the cost of holding money is sufficiently high, then the portfolio problem has an interior solution which uniquely determines the level of real balances carried into the beginning of the period. If, however, the cost of holding money is low enough then the solution occurs at the boundary where consumers carry enough money to purchase the entire amount of inventory. That is, there is zero liquidity value for real balances $z > z(q^w)$.

An interior solution to the consumer's problem is given by,

$$i = \mu(n)\theta_r \left[\frac{u'(q_r^*) - R}{(1-\theta_r)u'(q_r^*) + \theta_r R} \right] \tag{22}$$

If their is insufficient inventory, buyers purchase all inventory. A buyer's reaction function is given by,

$$q_r(q_w) = \begin{cases} q_r^* & \text{if} \quad q_r^* \leq q_w \\ \\ q_w & \text{if} \quad q_r^* > q_w \end{cases} \tag{23}$$

I now move to a middleman's inventory decision in the wholesale market. The terms of trade are similar to the pure credit economy; except now the expected surplus in the retail market is affected by consumers' real money balances. The amount of inventory purchased is given by,

$$\max_{q^w} (\mu(n)/n)(1-\theta_r)S_r(q^r) + Rq^w - c(q^w)$$

where the size of the surplus in the retail market $S_r(q^r) = u(q^r(q^w, m)) + -Rq^r(q^w, m)$ now depends on the portfolio choice of a consumer. Thus, the amount of inventory purchased in the wholesale market depends on consumers' portfolio choices made at the end of the CM. A middleman forms expectations about consumer's portfolio decisions which dictate the expected surplus in retail trades. Given these beliefs, a middleman optimally invests in inventory to maximize his period consumption.

I represent the middleman's problem graphically in Figure 10. A middleman weighs the cost of acquiring inventory against its expected value in the retail market. A middleman realizes positive marginal benefit from carrying extra inventory into the retail market up to some threshold $q^{-1}(z)$ which describes the amount of inventory that a consumer can purchase holding z real balances. Any additional inventory in excess of this threshold yields zero marginal benefit to a middleman. If this threshold is sufficiently high there is an interior solution and the middleman buys less inventory than a consumer is able to purchase with z real balances. If, however, the threshold is sufficiently low the middleman has a boundary solution where he buys exactly the amount of inventory that a consumer can purchase.

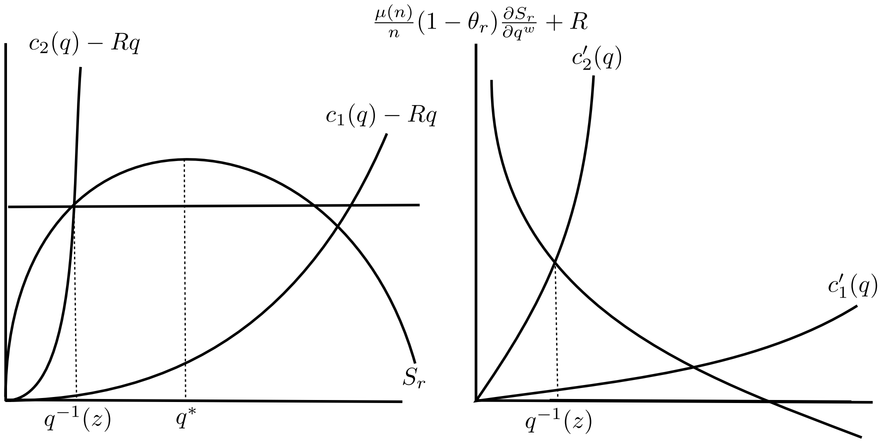

Figure 10: Middleman's Inventory Decision

An interior solution to the middleman's inventory problem is given by,

$$\frac{\mu(n)}{n}(1-\theta_r)\frac{\partial S_r(q_w^*)}{\partial q_w} + R = c'(q_w^*) \tag{24}$$

If a middleman anticipates that buyers carry to few real balances, then a middleman will purchase just as much inventory as it expects it can sell. The middleman's reaction function is given by,

$$q_w(q_r) = \begin{cases} q_w^* & \text{if} \quad q_w^* \leq q_r \\ \\ q_r & \text{if} \quad q_w^* > q_r \end{cases} \tag{25}$$

An equilibrium is defined as follows: (22),(23),(24),(25) determine the quantity traded for a given level of entry and (11) determines the level of entry for a given quantity traded. Notice that the ex-ante investment decisions by middlemen and consumers represented by (22)-(25) generate coordination failures that generate a continuum of equilibria indexed by $q \in [0, \min\{q_w^*, q_r^*\}]$. Figure 11 represents these equilibria for a given level of entry. The coincidence of reaction functions along the forty-five degree line constitute a continuum of equilibria enforced by self-fulfilling beliefs. Suppose that consumers anticipate middleman will carry q units of inventory and therefore hold $z(q)$ real money balances. Concurrently, middlemen anticipate consumers hold $z(q)$ real balances and response by investing in q units of inventory. Both agents beliefs are validated and an equilibrium obtains.

Relative to the pure credit economy, the amount of inventory can be no greater. The underin-

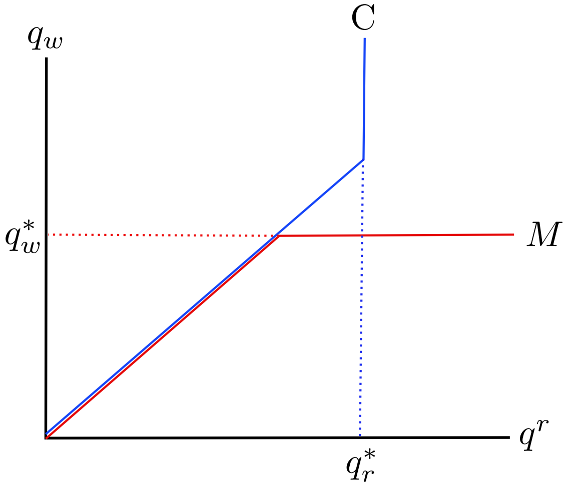

Figure 11: Equilibria Inventory

vestment problem is weakly worse. Weak in the sense that if consumers hold enough real balances, then the amount of inventory is the same as under the pure credit economy; however, if consumers hold too few real balances then there is more underinvestment in inventory. Making credit infeasible in the retail market (and so long as money is costly to hold) necessitates a weakly smaller surplus in the retail market. This decreases the value of holding inventory for a middleman.

Note the effect of nominal interest rates on the quantity of inventory. Conventionally, a higher nominal interest rate increases the opportunity cost of holding money which leads to fewer real balances and less trade. Consider, however, an equilibrium the consumer is at a boundary solution. In this case, the choice of real money balances is unaffected by a small change in the nominal interest rate. Even though the cost of holding money decreases, agents will not accumulate more money because they know such extra balances will be useless given that middlemen do not carry enough inventory. The quantity traded will only respond to the nominal interest rate along the set of interior solutions to the consumer's portfolio problem.

Money in retail transactions yields qualitatively different effects than under the pure credit economy. Consider the relationship between the measure of middlemen n and the amount of inventory purchased q^w. Suppose, initially, that consumer's portfolio decision has an interior solution and thus a middleman is at a boundary solution. Now suppose that more middlemen enter the market $\uparrow n$. This decreases the expected value of retail trade for middlemen resulting in a leftward shift of the inventory demand curve in Figure 10. Lower inventory demand reduces q^w and thus shifts the boundary condition for consumers to the left. Concurrently, greater entry increases the expected value of retail trade for consumers causing a rightward shift in the money demand curve in Figure 9. This causes the boundary condition for middlemen to shift right $\uparrow q^{-1}(z)$. If the increase in n is small, then the consumer is still at an interior solution, the middleman at a boundary solution, and the quantity of inventory *increases*. However, for a large increase in n, inventory demand shifts so far to the left that the consumer is at its boundary solution while the middleman is at an interior solution. This implies a *lower* amount of inventory.

PROPOSITION 6 *When money is essential in retail trades, the response of inventory to the measure of active middlemen is non-monotone. Define $q = \min\{q^r, q^w\}$ to be the quantity traded given by (22),(24). We have that*

$$\begin{aligned} \partial q/\partial n > 0 \quad & \textit{for} \quad (0, \bar{n}) \\ \partial q/\partial n < 0 \quad & \textit{for} \quad (\bar{n}, \infty) \end{aligned}$$

where $\bar{n}$ is such that $q^r = q^w$.

This is substantively different from the pure credit case due to the portfolio decision of consumers. For $(0, \bar{n})$ an increase in n incentivizes consumers to hold more money balances and middlemen rationally respond by increasing their purchase of inventory. For $(\bar{n}, \infty)$ an increase in n incentivizes middlemen to purchase few inventories and consumers rationally respond by holding fewer real balances.

An equilibrium in (26),(27),(23) is represented in Figure 12. Notice that the strategic complementarities between portfolio decisions and entry generate multiple equilibria. I denote the "high" equilibrium as (q_H, n_H) and the "low" equilibrium as (q_L, n_L). Both equilibria are supported by consistent and validated beliefs of agents. Consider the high equilibrium as an example. Suppose that middlemen anticipate that consumers will hold large real balances, and therefore anticipate a large surplus in retail trades. This incentivizes a large measure of entrants which increases the frequency of consumption opportunities making it advantageous for consumer's to hold large real balances, which supports firms' beliefs. Similarly, if firms believe consumers will hold few real balances, then entry is low, consumption opportunities are rare, and consumers hold few real bal-

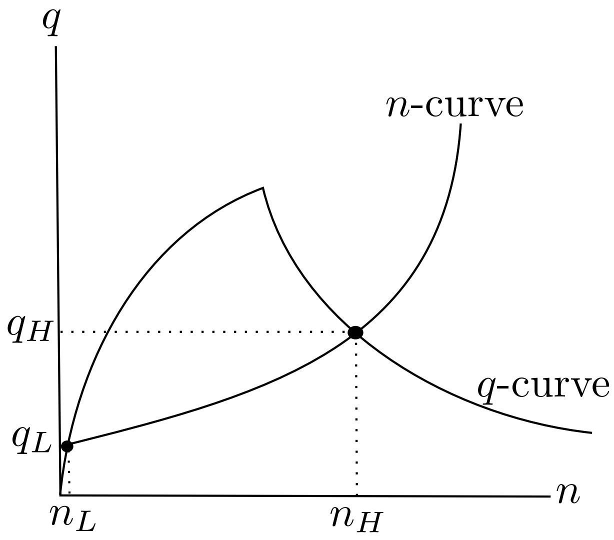

Figure 12: Equilibrium in (q^w, n)

ances which validates firms' beliefs. For the comparative statics that follow I focus on the high equilibrium.

PROPOSITION 7 *When money is essential in retail trades, the comparative statics in (q_H, n_H) depend on the location of the initial equilibrium. The comparative statics for an initial equilibrium $n_0 > \bar{n}$ are as follows: $\partial q_H/\partial\theta_r < 0, \partial n_H/\partial\theta_r < 0, \partial q_H/\partial\theta_w < 0, \partial n_H/\partial\theta_w > 0, \partial q_H/\partial i = 0, \partial n_H/\partial i = 0$. The comparative statics for an initial equilibrium $n_0 < \bar{n}$ are as follows: $\partial q_H/\partial\theta_r > 0, \partial n_H/\partial\theta_r > 0, \partial q_H/\partial\theta_w > 0, \partial n_H/\partial\theta_w > 0, \partial q_H/\partial i > 0, \partial n_H/\partial i > 0$.*

Increasing consumers' bargaining power $\uparrow \theta_r$ increases retail demand $\uparrow q^r$ and decreases inventory demand $\downarrow q^w$ for any given level of entry n. This results in a leftward shift of the q-curve shown in Figure 13.[7] Concurrently, the n-curve will rotate clockwise about q^w. The left diagram in Figure 15 shows the partial effect of $\uparrow \theta_r$ on the equilibrium due to the q-curve. The right diagram in Figure 13 shows the general equilibrium effects accounting for free entry via the n-curve.

The comparative statics depend on where the initial equilibrium is located. If initial equilibrium is at a level of entry $n > \bar{n}$ (where middlemen inventory demand is binding) then increasing consumers' bargaining power will result in fewer entrants and less quantity traded. Middlemen, facing a worse bargaining position, demand less inventory and consumers rationally respond by holding fewer real balances. If, however, the initial equilibrium is at some $n < \bar{n}$ (where consumers'

[7] It can also be shown analytically that $\partial\bar{n}/\partial\theta_r < 0$ which verifies the leftward shift of the q-curve.

demand is binding) then there will be more entrants and more quantity traded. Greater bargaining power incentivizes consumers to hold more real balances and middlemen rationally respond by purchasing more inventory. In the extreme case where $\theta_r = 1$, the q-curve shifts far to the left and has a horizontal portion corresponding to $q^w : c'(q^w) = R$ indicating that middlemen do not realize any value from holding inventory in the retail market. Concurrently, the n-curve rotates clockwise.

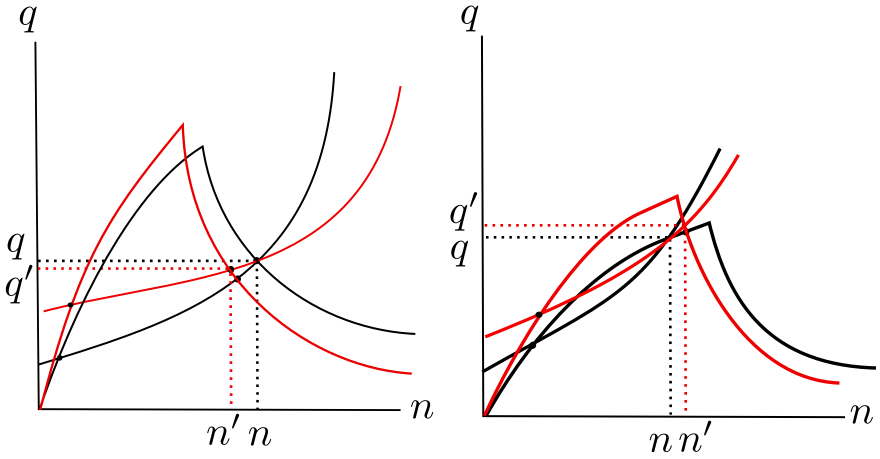

Figure 13: Increase in θ_r

Changes in bargaining power in the wholesale market have no effect on the q-curve but effect entry through the n-curve. More bargaining power for middlemen generates a larger expected surplus from entry which rotates the n-curve clockwise indicating more entry for any given level of trade. If the initial equilibrium is at $n > \bar{n}$ then $\partial n/\partial\theta_w > 0$ and $\partial q/\partial\theta_w < 0$. More entry induces congestion in the retail market resulting in downward movement along the inventory demand curve. If the initial equilibrium is $n < \bar{n}$ then $\partial n/\partial\theta_w > 0$ and $\partial q/\partial\theta_w > 0$. More entry increases the expected value of retail trade for consumers who respond by holding more real balances. These comparative statics are represented in Figure 14.

The above comparative statics suggest that the value of θ_w can determine where the initial equilibrium lies. If θ_w is large, middlemen will receive a large fraction of its expected surplus which incentivizes a large measure of entrants for any given quantity traded and the resulting equilibrium will be at $(q_H, n_H) : n_H > \bar{n}$. If, however, θ_w is small, then there will be few entrants and the equilibrium will be at $(q_H, n_H) : n_H < \bar{n}$.

Finally, I consider the effects of monetary policy. As suggested by Figure 9, there is a region over which monetary policy is ineffective. This occurs when inventory demand is a binding constraint for

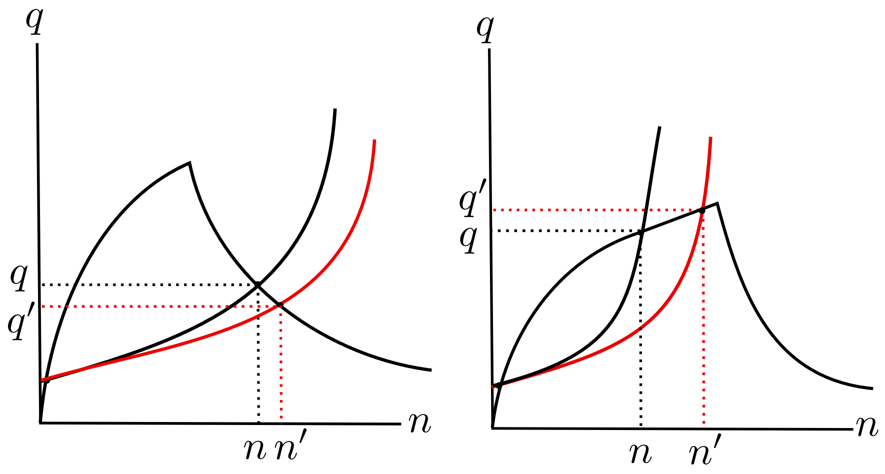

Figure 14: Increase in θ_w

the consumer. For any small change in the nominal interest rate, the quantity traded is unchanged because consumers realize zero liquidity value from holding extra money. Figure 15 shows that this is the case for all initial equilibria with $n > \bar{n}$. If the initial equilibrium is $n < \bar{n}$ then a lower nominal interest rate increases money demand and middlemen respond by holding greater inventory, and more trade in the retail market attracts new entrants. [8]

The efficacy of monetary policy may then crucially depend on how much bargaining power middlemen possess. If θ_w is large, then the high equilibrium will be such that monetary policy has no effect. The intuition is as follows. When middlemen receive a large share of future surpluses, there is a large measure of entry which increases competitive pressures in the wholesale market and results in less inventory acquisition. Since middlemen are purchasing too little inventory, consumers' will not realize the full value of their real balances and so rationally choose to hold only enough to purchase all inventory. The high equilibrium is thus characterized by too little inventory and consumers' facing a boundary solution which leaves monetary policy ineffective. Conversely, suppose that θ_w is small, so that there are relatively few entrants. Fewer entrants increases the probability of a retail trade which incentivizes middlemen to purchase more inventory. Consumers' are now able to realize the full return on their real balances since middlemen hold large inventories, and monetary policy is effective.

[8] There exists some threshold nominal interest rate $\bar{i}$ which defines the effective lower bound on interest rates, below which monetary policy is ineffective. This threshold is implicitly defined as follows: let (q_i, n_i) solve (11) and (24), then $\bar{i}$ is such that (22) holds at (q_i, n_i).

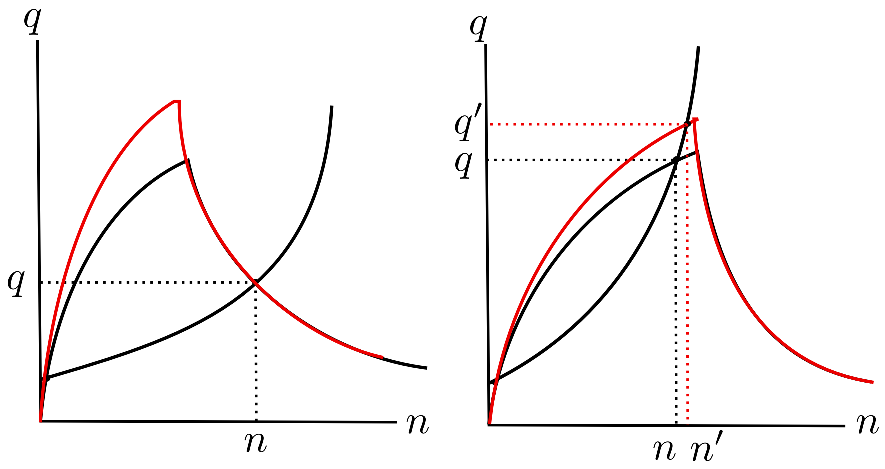

Figure 15: Decrease in i

10 Conclusion

The framework lends itself to developing an intermediation theory of the firm, first articulated by Coase (1937) and later refined by Spulber (1999), while including micro-foundations for the role of liquid assets. Stated simply, firms act as a conduit between suppliers and customers when the gains from intermediated trade are greater than the gains from direct trade. The conditions under which this happens are many, and always depend on the environment described by the modeler.

Presently, middlemen are merchants who buy and resell goods without engaging in any productive activity; while producers are a technology allowing for the manufacture of retail goods. The model can be amended so that middlemen more closely resemble firms in the conventional sense. Producers are reinterpreted as entrepreneurs who have an idea or ability to generate some input into a larger production process. Middlemen are reinterpreted as firms who purchase inputs from entrepreneurs, transform inputs into final consumption goods, and sell to consumers. The value-adding productive process employed by middlemen/firms can be formalized by positing a concave technology Q=G(q). Consumers then enjoy utility u(Q).

The reinterpreted framework places middlemen as the creators and operators of markets. They form bid and ask prices, conduct transactions, and allocate goods. The theory offers an explicit mechanism by which markets clear and equilibrium prices obtain rather than resorting to the theoretical construct of a Walrasian auctioneer.

It is worthwhile to consider alternative market structures while retaining middlemen as an explicit mechanism by which prices are set and quantities are determined. The obvious market structure to explore would be competitive price posting which allows one to price congestion in the market. This market structure may also be a more realistic representation of middlemen as market-makers rather than merchants. Watanabe (2017) considers the case of a monopolistic middleman who can choose whether to act as a merchant or a market-maker.

References

Gary Biglaiser. Middlemen as experts. *The RAND Journal of Economics*, 1993.

Xun Gong. Middlemen: Intensive and extensive margins with endogenous meeting technology. 2017.

Alok Johri and John Leach. Middlemen and the allocation of heterogeneous goods. *International Economic Review*, 2002.

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

Yiting Li. Middlemen and private information. *Journal of Monetary Economics*, 1998.

Yiting Li. Money and middlemen in an economy with private information. *Economic Inquiry*, 1999.

Adrian Masters. Unpleasent middlemen. *Journal of Economic Behavior and Organization*, 2008.

Ed Nosal, Yuet-Yee Wong, and Randall Wright. Buyers, sellers and middlemen: Variations in search theory. 2011.

Ed Nosal, Yuet-Yee Wong, and Randall Wright. More on middlemen: Equilibrium entry and efficiency in intermediated markets. 2014.

Guillaume Rocheteau and Randall Wright. Money in search equilibrium, in competitive equilibrium, and in competitive search equilibrium. *Econometrica*, 2005.

Ariel Rubinstein and Asher Wolinsky. Middlemen. *Quarterly Journal of Economics*, 102(3):581–593, 1987a.

Ariel Rubinstein and Asher Wolinsky. Middlemen. *The Quarterly Journal of Economics*, 1987b.

Shevchenko. Middlemen. *International Economic Review*, 2004.

Watanabe. A model of merchants. *Journal of Economic Theory*, 2010.

A Coexistence of Money and Credit in Retail Market

This section considers an environment where both money and credit are used in retail trades. The results from Section 4 and Section 6 are limiting cases of the environment described here.

Consumers can use two types of payments in retail trade: money and credit. In a fraction ω of retail trades, consumers' debt obligations are recorded and there exists perfect enforcement to guarantee repayment of the debt. In the remaining $1-\omega$ fraction of retail trades consumers are unmonitored precluding the use of credit so that only money can serve as payment.[9] The acceptability of credit is thus exogenous and random.

The expected value of entering the decentralized markets are given by the following value functions:

$$\begin{aligned}
&V_t^C = \mu(n)\left\{\omega[u(q^r)-b^r]+(1-\omega)[u(q^r)-\phi d^r]\right\}+W_t^C(m)\\
&V_{1,t}^M(\tilde{m})\frac{\gamma(n)}{n}\left\{\frac{\mu(n)}{n}\left[\omega(b^r-Rq^r)+(1-\omega)(\phi d^r-Rq^r)\right]+rq^w-b^w\right\}+W_t(0,0,\tilde{m})\\
&V_{2,t}^M(q,b,\tilde{m}) = \frac{\mu(n)}{n}\left\{\omega(b^r-Rq^r)+(1-\omega)(\phi d^r-Rq^r)\right\}+W_t^M(q,b,\tilde{m})\\
&V_t^P(\hat{m}) = \gamma(n)(-c(q^w)+b^w)+W_t^P(0,\hat{m})
\end{aligned}$$

I continue to settle the terms of trade by proportional bargaining. Buyers' portfolio choice now becomes,

$$\max_{q^r} -iz(q_r)+\mu(n)(1-\omega)\theta_r S_r(q_r)$$

$$s.t. \quad q_r \leq q_w$$

A buyer's reaction function is,

$$q_r(q_w) = \begin{cases} q_r^* & \text{if} \quad q_r^* \leq q_w \\ q_w & \text{if} \quad q_r^* > q_w \end{cases}$$

[9] In monitored matches consumers could use a mix of money and credit, but this arrangement is payoff-equivalent to using only credit.

where the interior solution is now given by,

$$i = \mu(n)(1-\omega)\theta_r \left[\frac{u'(q_r^*) - R}{(1-\theta_r)u'(q_r^*) + \theta_r R} \right]$$

A middleman's inventory choice is given by,

$$\max_{q_w} \frac{\mu(n)}{n}(1-\theta_r)\left\{\omega(u(q^r) - Rq^r) + (1-\omega)(u(q^r) - Rq^r)\right\} + Rq_w - c(q_w)$$

$$s.t. \quad q_w \leq q_r$$

A middleman's reaction function is,

$$q_w(q_r) = \begin{cases} q_w^* & \text{if} \quad q_w^* \leq q_r \\ q_r & \text{if} \quad q_w^* > q_r \end{cases}$$

where its interior solution is

$$\frac{\mu(n)}{n}(1-\theta_r)\frac{\partial S_r(q_w^*)}{\partial q_w} + R = c'(q_w^*)$$

where $S_r(q) = \omega(u(q^r) - Rq^r) + (1-\omega)(u(q^r) - Rq^r)$ is the expected surplus in retail trades for a given level of credit availability ω.