LIQUIDITY:
A NEW MONETARIST PERSPECTIVE*

Ricardo Lagos
NYU
Guillaume Rocheteau
UC-Irvine
Randall Wright
UW-Madison, FRB Minneapolis, FRB Chicago, and NBER

August 19, 2015

Abstract

This essay surveys the New Monetarist approach to liquidity. Work in this literature strives for empirical and policy relevance, plus rigorous foundations. Questions include: What is liquidity? Is money essential in achieving desirable outcomes? Which objects can or should serve in this capacity? When can asset prices differ from fundamentals? What are the functions of commitment and collateral in credit markets? How does money interact with credit and intermediation? How can and should monetary policy do? The research summarized emphasizes the micro structure of frictional transactions, and studies how institutions like monetary exchange, credit arrangements or intermediation facilitate the process.

*A previous version of this essay circulated as “The Art of Monetary Theory: A New Monetarist Perspective” (to pay homage to Clarida et al. 1999). For comments we thank the editor and four referees. Particularly useful input was also provided by David Andolfatto, Ed Green, Ricardo Cavalcanti, Mei Dong, John Duffy, Pedro Gomis-Porqueras, Ian King, Yiting Li, Narayana Kocherlakota, Andy Postlewaite, Daniella Puzzello, Mario Silva, Alberto Trejos, Neil Wallace, Steve Williamson and Sylvia Xiao. Parts of the essay were delivered by Wright as the 2011 Toulouse Lectures over a week at the Toulouse School of Economics, and he thanks them for their generous hospitality. Wright also acknowledges the NSF and the Ray Zemon Chair in Liquid Assets at the Wisconsin School of Business for research support. Lagos acknowledges the C.V. Starr Center for Applied Economics at NYU as well as the hospitality of the Cowles Foundation for Research in Economics at Yale University, the Federal Reserve Bank of Minneapolis and the Federal Reserve Bank of St. Louis. The usual disclaimers apply.
This, as I see it, is really the central issue in the pure theory of money. Either we have to give an explanation of the fact that people do hold money when rates of interest are positive, or we have to evade the difficulty somehow... The great evaders ... would have put it down to “frictions,” and since there was no adequate place for frictions in the rest of their economic theory, a theory of money based on frictions did not seem to them a promising field for economic analysis. This is where I disagree. I think we have to look frictions in the face, and see if they are really so refractory after all. Hicks (1935)

Progress can be made in monetary theory and policy analysis only by modeling monetary arrangements explicitly. Kareken and Wallace (1980)

1 Introduction

Over the past 25 years a new approach has been developed to study monetary theory and policy, and, more broadly, to study liquidity. This approach sometimes goes by the name New Monetarist Economics. Research in the area lies at an interface between macro and micro – it is meant to be empirically and policy relevant, but it also strives for theoretical rigor and logical consistency. Of course most economic research tries to be rigorous and consistent, but it seems accurate to say the body of work we call New Monetarism is particularly concerned with microfoundations. Recently, however, the field has become increasingly policy oriented as the theories have matured, and as recent events have put monetary matters front and center, including those related to interest rates, banking, credit conditions, financial markets and liquidity.¹

Papers in the area are diverse, yet share a set of principles and methods. Common themes include: What exactly is liquidity? What is the role of money and is it essential for improving the allocation of resources? Which objects will or should

¹Williamson and Wright (2010a) discuss the New Monetarist label and compare the approach to Old Monetarist and New or Old Keynesian economics. Briefly, at least some people contributing to the literature surveyed here find attractive many (not all) ideas in Old Monetarism, represented by Friedman (1960,1968,1969) and his followers. There is also opposition to the New Keynesians (see fn. 6 for citations) due to disagreements over what constitutes solid microfoundations. While there is a broad consensus on the attractiveness of microfounded dynamic models, in general, there are differences about which frictions matter most. We tend to focus on limited commitment, imperfect monitoring, private information and difficulties in coordinating trade (search frictions). Keynesians focus on nominal price rigidity as the critical, if not the exclusive, distortion relevant for macroeconomic analysis. We hope that explaining the areas of disagreement is constructive, the way it was healthy for Old Monetarists and Keynesians to debate the issues.
play this role in equilibrium or optimal arrangements? Why is intrinsically worthless currency valued, or, more generally, why do asset prices differ from their fundamental values? How does credit work absent commitment? What is the role of collateral? Under what conditions can money and credit coexist? What are the functions of intermediation? What are the effects of inflation? What can monetary policy achieve and how should it be conducted? As will become clear, the research concerns much more than just pricing currency. It concerns trying to understand the process of exchange in the presence of explicit frictions, and how this process might be facilitated by institutions including money, credit, collateral and intermediation.

A characteristic of the literature discussed below is that it models the transactions process explicitly, in the sense that agents trade with each other. That is not true in GE (general equilibrium) theory, where they merely slide along budget lines. In Arrow-Debreu, agents are endowed with a vector $\bar{x}$, and choose another $x$ subject only to $px \leq p\bar{x}$, with $p$ taken as given; how they get from $\bar{x}$ to $x$ is not modeled.\(^2\) The models below incorporate frictions that hinder interactions between agents and then analyze how institutions ameliorate these frictions. Since money is one such institution, maybe the most elemental, it is natural to start there but the work does not stop there. We are interested in any institution whose raison d’être is the facilitation of transactions, and for this, one needs relatively explicit descriptions of the trading process.\(^3\)

\(^2\)Earlier research surveyed by Ostroy and Starr (1990) asked some of the right questions, but did not resolve all the issues. Also related is work trying to use Shapley and Shubik’s (1969) market games as a foundation for monetary economics, e.g. Hayashi and Matsui (1996).

\(^3\)This is a venerable concern. On “the old conundrum” of fiat money, Hahn (1987) says: “At a common-sense level almost everyone has an answer to this, and old-fashioned textbooks used to embroider on some of the banalities at great length. But common sense is, of course, no substitute for thought and certainly not for theory. In particular, most of the models of an economy which we have, and I am thinking here of many besides those of Arrow and Debreu, have no formal account for the exchange process.” Clower (1970) similarly says: “conventional value theory is essentially a device for logical analysis of virtual trades in a world where individual economic activities are costlessly coordinated by a central market authority. It has nothing whatever to say about delivery and payment arrangements, about the timing or frequency of market transactions, about search, bargaining, information and other trading costs, or about countless other commonplace features of real-world trading processes.” On middlemen, Rubinstein and Wolinsky (1987) say: “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.”
Research discussed below borrows from GE, and some models adopt the abstraction of competitive markets, when convenient, but others rely heavily on search theory, which is all about agents trading with each other. They also use game theory since the issues are inherently strategic, e.g., what one accepts as a medium of exchange or collateral might depend on what others accept. At the intersection of search and game theory, the papers often use bargaining, which leads to insights one might miss if exclusively wed to a Walrasian auctioneer. Other ways to determine the terms of trade are studied, too, including price taking, posting and more abstract mechanisms. Key frictions include spatial and temporal separation, limited commitment and imperfect information. These can make arrangements involving money socially beneficial, which is not the case in models with CIA (cash-in-advance) or MUF (money-in-utility-function) assumptions. Such reduced-form devices presumably stand in for the idea that monetary exchange helps overcome some difficulties – but then, one might ask, why not model that?

In CIA models, having to use cash hurts; in sticky-price models, nothing but problems arise from having agents set prices in dollars and making it difficult or costly to change. If money were really such a hindrance, how did it survive all these centuries? The work reviewed here tries to get the relevant phenomena, like monetary or credit arrangements, to arise endogenously, as beneficial institutions. This can give different answers than reduced-form models, and allows additional questions. How can one purport to understand financial crises or banking problems using theories with no essential role for payment or settlement systems in the first place? How can one hope to assess the effects of inflation – a tax on the use of money – using theories that do not incorporate the frictions that money is meant to remedy? We do not claim the papers discussed below have definitive answers to all these questions, but contend they provide useful ways to think about them.

To hint at where we are going, consider these well-known examples of work that shares our stance on avoiding shortcuts to model monetary phenomena. There is a large body of work using OLG (overlapping-generations) economies, with major contributions by Samuelson (1958), Lucas (1972) and Wallace (1980). Since Kiyotaki and Wright (1989), more monetary economists employ search theory (Jones 1976,
Oh 1989 and Iwai 1996 are other early attempts at search-based models of money. However, as discussed below, spatial separation per se is not crucial, as clarified by Kocherlakota (1998), in line with earlier work by Ostroy (1973) and Townsend (1987b). On banking, we discuss research drawing on Diamond and Dybvig (1983), Diamond (1984) and Williamson (1986, 1987). On intermediation, related contributions include Rubinstein and Wolinsky (1987) and Duffie et al. (2005). On secured credit, the models share themes with Kiyotaki and Moore (1997, 2005), while on unsecured credit, they make use of Kehoe and Levine (1993) and Alvarez and Jermann (2000).  

By way of preview, we start with the first-generation search models of money by Kiyotaki and Wright (1989, 1993), Aiyagari and Wallace (1991, 1992) and others, to illustrate tradeoffs between asset returns and acceptability, and to show how economies where liquidity plays a role are prone to multiplicity and volatility. We also present elementary versions of Kocherlakota’s (1998) results on the essentiality of money, and Cavalcanti and Wallace’s (1999a) analysis of inside and outside money. We then move to the second-generation models by Shi (1995), Trejos and Wright (1995) and others, with divisible goods. This permits further exploration of the efficiency of monetary exchange and the idea that liquidity considerations lead to multiplicity and volatility. We also develop a connection with the above-mentioned models of intermediation.

We then move to divisible assets, as in computational work by Molico (2006) and others, or the more analytically tractable approaches in Shi (1997a) and Lagos and Wright (2005). These models are easier to integrate with mainstream macro and allow us to examine many standard issues in a new light. We also discuss the effects of monetary policy on labor markets, the interaction between money and other assets in facilitating exchange, and the theme that observations that seem anomalous from the perspective of standard financial economics can emerge naturally when there are trading frictions. Also, just as some papers relax the assumption of indivisible

---

4Further on method, our preference for modeling monetary, credit and other such arrangements explicitly is related to the Lucas (1976) critique, ideas espoused in Townsend (1987a, 1988), and the Wallace (1998) dictum that says: “Money should not be a primitive in monetary theory in the same way that firm should not be a primitive in industrial organization theory or bond a primitive in finance theory.”
assets in early search-based monetary theory, we show how Lagos and Rocheteau (2009) relax it in models of OTC (over-the-counter) asset markets along the lines of Duffie et al. (2005). We also discuss information-based models.

Before getting into theory, let us clarify a few terms and put them in historical perspective. It is commonly understood that double-coincidence problems plague direct barter: in a bilateral meeting between individuals \( i \) and \( j \), it would be a \textit{coincidence} if \( i \) liked \( j \)'s good, and a \textit{double coincidence} if \( j \) also liked \( i \)'s good.\(^5\) Reference is often made to a \textit{medium of exchange}, and on this it is hard to improve on Wicksell (1911): “an object which is taken in exchange, not for its own account, i.e. not to be consumed by the receiver or to be employed in technical production, but to be exchanged for something else within a longer or shorter period of time.” A medium of exchange that is also a consumption or production good is \textit{commodity money}. By contrast, Wallace (1980) defines \textit{fiat money} as a medium of exchange that is intrinsically useless (neither a consumption good nor a productive input) and inconvertible (not a claim on consumption or production goods). This usefully delineates a pure case, although assets other than currency can convey \textit{moneyness} – i.e., can be more or less easy to use in transactions.

A challenge in monetary economics is to describe environments where an institution like money is \textit{essential}. As introduced by Hahn (1973), essentiality means welfare is higher, or the set of incentive-feasible allocations is larger, with money than without it. In some applications, it is fairly evident that one should want to use models where money is essential – e.g., again, to understand the cost of the inflation tax, one might want to first understand the benefits of monetary exchange, and the fact that money is so prevalent over time suggests that as an institution it must be welfare improving. That this is nontrivial is evidenced by the fact that money is not essential in standard theories. As Debreu (1959) says about his, an “important and difficult question ... not answered by the approach taken here: the integration of money in the theory of value.” An objective here is to clarify what

\(^5\)Jevons’ (1875) often gets credit, but other discussions of this idea, in reverse historical order, include von Mises (1912), Wicksell (1911), Menger (1892) and Smith (1776). Sargent and Velde (2003) argue that the first to see how money helps with a double-coincidence problem was the Roman jurist Paulus in the 2nd century, despite Schumpeter’s (1954) claim for Aristotle.
ingredients are relevant for getting liquidity considerations into rigorous theory, and
to show how economics gets more interesting when they are there.\footnote{Related surveys or discussions include Wallace (2001,2010), Wright (2005), Shi (2006), Lagos (2008), Williamson and Wright (2010a,b) and Nosal and Rocheteau (2011). By way of comparison, in this essay we: (i) discuss liquidity generally with money a special case; (ii) connect more to finance and labor; (iii) provide simplified versions of difficult material not found elsewhere; (iv) highlight a few quantitative results; and (v) cover work over the past 5 years. Surveys of the very different New Keynesian approach include Clarida et al. (1999) and Woodford (2003). To be clear, while there are disagreements, it is certainly desirable that different practices are allowed to flourish. Also, it is hard not to be impressed by Keynesian success with policy makers and in the media; the more modest goal here is to communicate to scholars that there is an alternative.}

## 2 Commodities as Money

We begin with Kiyotaki and Wright (1989). The goal is to derive equilibrium transactions patterns endogenously and see if they resemble trading arrangements in actual economies, in a stylized sense, such as certain commodities acting as media of exchange, or certain agents playing the role of middlemen. While abstract relative to much of what follows, this goal still seems relevant, and a discussion seems appropriate, because this setting provides a rudimentary and early example of the methods used in the literature.\footnote{Without much loss of continuity, it is possible to skip to Section 6 in order to reach the frontier more quickie, or to skip to Section 3 for a model that came later but is easier than this one.}

Time is discrete and continues forever. There is a $[0, 1]$ continuum of infinitely-lived agents that meet bilaterally at random. To introduce gains from trade, assume they have specialized tastes and technologies: there are $N$ types of agents and $N$ goods, where type $j$ agents consume good $j$ but produce good $j + 1$ modulo $N$ (i.e., type $N$ agents produce good $N + 1 = 1$). For now, the fraction of type $j$ is $n_j = 1/N$, and we set $N = 3$. Although we usually call the agents consumers, obviously similar considerations are relevant for producers. Instead of saying individual $i$ produces good $i + 1$ but wants to consume $i$, it is a relabeling to say firm $i$ uses $i$ to produce $i + 1$, generating the same motives for and difficulties with trade. However, although these difficulties hinder barter, the double-coincidence problem per se does nothing to rule out credit. For that we need a lack of enforcement/commitment plus imperfect information, as assumed here and discussed in detail in Section 3.

Goods are indivisible. They are storable only 1 unit at a time. Let $\rho_j$ be the
return on good \( j \) – i.e., the flow utility agents get from a unit of it in inventory. One can interpret \( \rho_j > 0 \) as a dividend, or fruit from a Lucas (1978) tree, and \( \rho_j < 0 \) as a storage cost. As discussed in Nosal and Rocheteau (2011, Chapter 5), it is a venerable idea that the intrinsic properties of objects influence which can or should serve as media of exchange, and storability is the property in focus here. Type \( j \) agents also get utility \( u > 0 \) by consuming good \( j \), and then produce a new unit of good \( j + 1 \) at cost \( c = 0 \). Type \( j \) agents always accept good \( j \) in trade, and immediately consume, given \( |\rho_j| \) is not too big (more on this below).

The following aspect of strategies is to be determined: Will type \( j \) agents trade good \( j + 1 \) for good \( j + 2 \) in an attempt to facilitate acquisition of good \( j \)? Or will they hold onto good \( j + 1 \) until trading directly for good \( j \)? If \( \tau_j \) is the probability type \( j \) agents trade good \( j + 1 \) for good \( j + 2 \), a symmetric, stationary, strategy profile is a vector \( \tau = (\tau_1, \tau_2, \tau_3) \). If \( \tau_j > 0 \), type \( j \) agents use good \( j + 2 \) as a medium of exchange. Also to be determined is the distribution of inventories. Since type \( j \) agents consume good \( j \), they always have either good \( j + 1 \) or \( j + 2 \). Hence, \( m = (m_1, m_2, m_3) \) gives the distribution, where \( m_j \) is the proportion of type \( j \) agents with good \( j + 1 \). The probability type \( i \) agents meet type \( j \) agents with good \( j + 1 \) each period is \( \alpha n_j m_j \), where \( \alpha \) is the probability of meeting anyone and \( n_j = 1/N \).

The Appendix derives the SS (steady state) condition for each \( j \),

\[
m_j n_{j+1} m_{j+1} \tau_j = (1 - m_j) n_{j+2} m_{j+2}. \tag{1}
\]

To describe payoffs, let \( r \) be the rate of time preference and \( V = (V_{ij}) \), where \( V_{ij} \) the value function of type \( i \) holding good \( j \). For notational convenience, let the utility from dividends be realized next period. Then, for type 1,

\[
rV_{12} = \rho_2 + \alpha n_2 (1 - m_2) u + \alpha n_3 m_3 \tau_3 u + \alpha n_2 m_2 \tau_1 (V_{13} - V_{12}) \tag{2}
\]

\[
rV_{13} = \rho_3 + \alpha n_3 m_3 (u + V_{12} - V_{13}), \tag{3}
\]

which are standard DP (dynamic programming) equations; the Appendix provides more details for readers less familiar with the methods. The BR (best response) conditions are: \( \tau_1 = 1 \) if \( V_{13} > V_{12} \); \( \tau_1 = 0 \) if \( V_{13} < V_{12} \); and \( \tau_1 = [0, 1] \) if \( V_{13} = V_{12} \). Calculation implies \( V_{13} - V_{12} \) takes the same sign as

\[
\Delta_1 \equiv \rho_3 - \rho_2 + \alpha [n_3 m_3 (1 - \tau_3) - n_2 (1 - m_2)] u. \tag{4}
\]
In (4), \( \rho_3 - \rho_2 \) is the return differential from holding good 3 rather than good 2. If returns were all agents valued, this would be the sole factor determining \( \tau_1 \). But the other term is the difference in the probability of acquiring 1’s desired good when holding good 3 rather than 2, or the liquidity differential. Whether type 1 agents should opt for indirect exchange, swapping good 2 for 3 whenever they can, involves comparing return and liquidity differentials. This reduces the BR condition for \( \tau_1 \) to a check on the sign of \( \Delta_1 \). Indeed, for any \( \phi \):

\[
\tau_j = \begin{cases} 
1 & \text{if } \Delta_j > 0 \\
[0, 1] & \text{if } \Delta_j = 0 \\
0 & \text{if } \Delta_j < 0
\end{cases}
\]  

(5)

A stationary, symmetric equilibrium is a list \( \langle V, m, \tau \rangle \) satisfying the DP, SS and BR conditions. There are 8 candidate equilibria in pure strategies, and for each such \( \tau \), one can solve for \( m \), and use (5) to determine the parameters for which \( \tau \) is a BR to itself (see the Appendix).

To present the results, assume \( \rho_1, \rho_2 > 0 = \rho_3 \), so we can display outcomes in the positive quadrant of \( (\rho_1, \rho_2) \) space. Figure 1 shows different regions labeled by \( \tau \) to indicate which equilibria exist. There are two cases, Model A or B, distinguished by \( \rho_1 > \rho_2 \) or \( \rho_2 > \rho_1 \). In Figure 1, Model A corresponds to the region below the 45° line, where there are two possibilities: if \( \rho_2 \geq \hat{\rho}_2 \) the unique outcome is \( \tau = (0, 1, 0) \); and if \( \rho_2 \leq \hat{\rho}_2 \) it is \( \tau = (1, 1, 0) \). To understand this, note that for type 1 good 3 is more liquid than good 2. It is more liquid since type 3 agents accept good 3 but not good 2, and type 3 agents always have what type 1 wants. In contrast, type 2 agents always accept good 2 but only have good 1 with probability \( m_2 = 1/2 \). Hence, good 3 allows type 1 agents to consume sooner. If \( \rho_2 > \hat{\rho}_2 \) this liquidity factor does not compensate for a lower return; if \( \rho_2 < \hat{\rho}_2 \) it does. The reason type 1 is pivotal is this: for type 2 agents, trading good 3 for good 1 enhances both liquidity and return, as does holding onto good 1 for type 3. Hence, only type 1 agents have a tradeoff.

In Kiyotaki and Wright (1989), Models A and B both have \( \rho_1 > \rho_2 \), but in Model B type \( j \) produces good \( j - 1 \) instead of \( j + 1 \). Here we keep production fixed and distinguish A or B by \( \rho_1 > \rho_2 \) or \( \rho_2 > \rho_1 \), which is equivalent but easier. To see how they differ, consider Model A with \( \rho_1 > \rho_2 \). Then at least myopically it looks like a bad idea for type 1 to set \( \tau_1 = 1 \), because trading good 2 for 3 lowers his return. Similarly, it looks like a bad idea for type 3 to set \( \tau_3 = 1 \), and a good idea for type 2 to set \( \tau_2 = 1 \). Hence, exactly one type is predisposed to use indirect trade based on fundamentals. In Model B, types 2 and 3 are both so predisposed.
In Model A, \( \tau = (0, 1, 0) \) is called the fundamental equilibrium. It features good 1 as the universally-accepted commodity money, and has type 2 agents acting as middlemen by acquiring good 1 from its producers and delivering it to its consumers. While this is a natural outcome, if \( \rho_2 < \bar{\rho}_2 \) we instead get \( \tau = (1, 1, 0) \), called a speculative equilibrium. This outcome has type 1 agents trading good 2 for the lower-return good 3 to improve their liquidity position, and both good 1 and 3 are used for indirect exchange. Theory delivers cutoffs for type 1 to sacrifice return for liquidity, but there is a gap: for \( \hat{\rho}_2 > \rho_2 > \bar{\rho}_2 \) there is no stationary, symmetric equilibrium in pure-strategies. Kehoe et al. (1993) show there is one in mixed-strategies, where type 1 agents accept good 3 with probability \( \tau^* \in (0, 1) \). They also show there can be multiple stationary mixed-strategy equilibria, but the set of such equilibria is finite. Whenever \( \hat{\rho}_2 > \rho_2 > \bar{\rho}_2 \) they also construct nonstationary equilibria with \( \tau^* \) cycling over time – an early (perhaps the first?) example of production and exchange fluctuating as a self-fulfilling prophecy due to liquidity considerations.\(^9\)

In Model B, above the 45° line, there is always an equilibrium with \( \tau = (0, 1, 1) \). This is the fundamental equilibrium for Model B, where type 1 agents hang on to good 2, which now has the highest return, while types 2 and 3 opt for indirect

\(^9\)Trachter and Oberfield (2012) show the set of dynamic equilibria shrinks as the length of the period gets small in a version of the model. Still, a recurrent theme below is that economies where liquidity plays a role are generally prone to multiplicity and volatility.
exchange, with goods 1 and 2 serving as money. For some parameters, there coexists an equilibrium with \( \tau = (1, 1, 0) \), the speculative equilibrium for this specification, where good 2 is not universally accepted even though it now has the best return. The coexistence of equilibria with different transactions patterns and liquidity properties shows that these are not necessarily pinned down by fundamentals.

This is the baseline model. In an extension, Aiyagari and Wallace (1991) allow \( N \) types and \( N \) goods, and prove existence of an equilibrium where the highest-return good is universally accepted (but there can also exist equilibria where this good is not universally accepted). This is nontrivial because standard fixed-point theorems do not guarantee equilibrium with a particular exchange pattern (one of Hahn’s 1965 problems; see also Zhu 2003,2005). An extension by Kiyotaki and Wright (1989) and Aiyagari and Wallace (1992) is to add fiat currency. We postpone discussion of this, but mention that it provides one way to see equilibria are not generally efficient: for some parameters, equilibria with valued fiat money exist and dominate other equilibria. In terms of comparing commodity-money equilibria when they coexist, it may seem better to use the highest-return object as money, but some agents may prefer to have other objects so used, like those who produce these objects, reminiscent of the bimetalism debates (e.g., see Friedman 1992).

To study how we get to equilibrium, several papers use evolutionary dynamics.\(^{10}\) Wright (1995) has a general population \( n = (n_1, n_2, n_3) \), and in one application agents can choose their type. This can be interpreted as choosing preferences, or technologies, or as an evolutionary process where types with higher payoffs increase in numbers due to reproduction or imitation. In Model A, with \( n_t \) evolving according to standard Darwinian dynamics, for any initial \( n_0 \), and any initial equilibrium if \( n_0 \) admits multiplicity, \( n_t \rightarrow n_\infty \) where at \( n_\infty \) the unique equilibrium is speculative. Intuitively, with fundamental strategies type 3 agents get the highest payoff, since they produce a good with the best return and highest liquidity. Ergo, \( n_3 \) increases.

\(^{10}\)These include Matsuyama et al. (1993), Wright (1995), Luo (1999) and Sethi (1999). Relatedly, Marimon et al. (1990) and Başçı (1999) ask if artificially-intelligent agents can learn to play equilibrium in the model. There are also studies in the lab. In these experiments, Brown (1996) and Duffy and Ochs (1999) find subjects have little problem finding the fundamental equilibrium, but can be reluctant to adopt speculative strategies. Duffy (2001) shows they can learn to do so. Duffy and Ochs (2002) also experiment with versions including fiat currency.
and as type 1 agents interact with type 3 more often, they are more inclined to sacrifice return for liquidity.

Motivated in part by criticism of random matching (e.g., Howitt 2005 or Prescott 2005), Corbae et al. (2003) redo the model with directed search. Generalizing Gale and Shapley (1962), at each \( t \) the population partitions into subsets containing at most two agents such that there are no profitable deviations in trade or trading partners for any individual or pair. For Model A with \( n_i = 1/3 \), with directed search the fundamental outcome \( \tau = (0,1,0) \) is the unique equilibrium in a certain class. On the equilibrium path, starting at \( m = (1,1,1) \), type 2 trade with type 3 while type 1 sit out. Next period, at \( m = (1,0,1) \), type 2 trade with type 1 while type 3 sit out, putting us back at \( m = (1,1,1) \). Different from random matching, there is a unique outcome and it features good 1 as money. Heuristically, with random search, in speculative equilibrium type 1 cares about the chance of meeting type 3 with good 1; with directed search, chance is not a factor because the endogenous transaction pattern is deterministic. Indeed, type 2 agents cater to type 1 agents’ needs by acting as middlemen, delivering consumption every second period. Hence, one might say that some randomness is needed to make operative type 1’s precautionary demand for liquidity, as seems natural.\(^{11}\)

As a final application, Renero (1998,1999) shows that equilibria where all agents randomize exist for many parameter values, and in such equilibria goods with lower \( \rho \) have higher acceptability. Intuitively, to make agents indifferent between lower- and higher-return objects, the former must be more liquid. This outcome, which can be socially desirable, is related to Gresham’s Law, and more generally captures rigorously a robust idea: Abstracting from risk, for the sake of illustration, whenever agents are indifferent between two assets (e.g., savings and checking deposits), as they must be if they are willing to hold both, the one with a lower return must be more liquid.\(^{12}\)

\(^{11}\)However, these results have only been established for Model A and \( n_i = 1/3 \) fixed exogenously; it is not known what happens in Model B, for \( n_i \neq 1/3 \) or for \( n \) determined endogenously, so there is still work to be done on this model.

\(^{12}\)While this should suffice to illustrate how an early formalization of the liquidity concept works, additional results are available – e.g., Camera (2001) discusses intermediation in more detail, while Cuadras-Morató (1994) and Li (1995) incorporate recognizability considerations in versions with private information. We return to these topics in Sections 5 and 10.
3 Assets as Money

Adding other assets allows us to illustrate additional results: (i) assets can facilitate intertemporal exchange; (ii) this may be true for fiat currency, an asset with a 0 return, or even for those with negative returns; (iii) for money to be essential, necessary conditions include limited commitment and imperfect information; (iv) the value of fiat money is both tenuous and robust; and (v) whether assets circulate as media of exchange may not be pinned down by primitives.\footnote{The setup here follows Kiyotaki and Wright (1991,1993), simplified it in various ways. In the original model, as in Diamond (1982), agents go to one “island” to produce and another to trade; here they produce on the spot. Also, in early versions, agents consume all goods, but like some more than others and have to choose which to accept; here that choice is made obvious. Also, in early formulations agents with assets could not produce, and so had to use money even in double-coincidence meetings; that is relaxed in this presentation, following Siandra (1990).}

Goods are nonstorable and produced on the spot for immediate consumption, at cost $c > 0$. Hence, they cannot be re-traded (one might think of them as services). Agents still specialize, but now, when $i$ and $j$ meet, the probability of a double coincidence is $\delta$, the probability of a single coincidence where $i$ likes $j$’s output is $\sigma$, and symmetrically the probability of a single coincidence where $j$ likes $i$’s output is also $\sigma$. So the double-coincidence arrival rate is $\alpha \delta$, with $\alpha$ and $\delta$ capturing search and matching, respectively, and the equations below holding in discrete time or continuous time with $\alpha$ interpreted as a Poisson parameter. Any good that $i$ likes gives him the same utility $u > c$; all others give him 0. There is a storable asset that no one consumes but yields a flow utility $\rho$. If $\rho = 0$ it is fiat currency. For now, assume that agents neither dispose of nor hoard assets (we check this below).

Let $A \in [0, 1]$ be the fixed asset supply, and for now continue to assume assets are indivisible and agents can hold at most 1 unit, $a \in \{0, 1\}$. To begin, however, let $A = 0$, so that if credit is not incentive feasible barter is the only option. The flow barter payoff is $rV^B = \alpha \delta (u - c)$. Given $\delta > 0$, this beats autarky: $V^B > V^A = 0$. But given $\sigma > 0$, it does not do all that well, because in some meetings $i$ wants to trade but $j$ will not oblige, which is bad for everyone ex ante. Suppose we try to institute a credit system, where agent $i$ produces for $j$ whenever $j$ likes $i$’s output. This is credit because agents produce while receiving nothing by way of quid pro quo, with the understanding – call it a promise – that someone will do the same for
them in the future. The flow payoff to this arrangement is

\[ rV^C = \alpha \delta (u - c) + \alpha \sigma u - \alpha \sigma c = \alpha (\delta + \sigma) (u - c). \]

If \( \sigma > 0 \) then \( V^C > V^B \). If agents could commit, therefore, they would promise to abide by this arrangement, and this maximizes ex ante utility conditional on the matching process. However, if they cannot commit, we must check said promise is credible. This entails an incentive condition, IC, when an agent is supposed to produce in a single-coincidence meeting, and of course it depends on the consequences of deviating. If \( V^D \) is the deviation payoff, IC is

\[ -c + V^C \geq \mu V^D + (1 - \mu) V^C, \quad (6) \]

where \( \mu \) is the probability deviators are caught and punished. Thus, \( \mu < 1 \) captures imperfect monitoring, or record keeping, so that deviations are only probabilistically detected and communicated to the population at large. The best punishment is the harshest, which is banishment to autarky, but that may not be feasible, depending on details, so we take the best punishment to be a loss of future credit (these are of course the same if \( \delta = 0 \)). Then \( V^D = V^B \), in which case (6) holds iff

\[ r \leq \hat{r}_C \equiv \mu \alpha \sigma (u - c)/c. \quad (7) \]

As is standard, low \( r \) is necessary for cooperative behavior. Or, one can alternatively say that \( \mu \) must not be too small, meaning the monitoring/communication technology must be sufficiently good. If credit is not viable, consider monetizing exchange. Let \( V_a \) be the value function for agents with \( a \in \{0, 1\} \), and call those with \( a = 1 \) (\( a = 0 \)) buyers (sellers). Then

\[ rV_0 = \alpha \delta (u - c) + \alpha \sigma A \tau_0 \tau_1 (V_1 - V_0 - c) \quad (8) \]

\[ rV_1 = \alpha \delta (u - c) + \alpha \sigma (1 - A) \tau_0 \tau_1 (u + V_0 - V_1) + \rho, \quad (9) \]

where \( \tau_0 \) is the probability sellers are willing to produce for assets while \( \tau_1 \) is the probability buyers are willing to give up assets to consume, and we include \( \rho \) so the equations also apply to real assets. If \( \Delta = V_1 - V_0 \), the BR conditions are:

\[ \tau_0 = \begin{cases} 
1 & \text{if } \Delta > c \\
[0, 1] & \text{if } \Delta = c \\
0 & \text{if } \Delta < c
\end{cases} \quad \text{and} \quad \tau_1 = \begin{cases} 
1 & \text{if } u > \Delta \\
[0, 1] & \text{if } u = \Delta \\
0 & \text{if } u < \Delta
\end{cases} \quad (10) \]
Letting \( \mathbf{V} = (V_0, V_1) \) and \( \mathbf{\tau} = (\tau_0, \tau_1) \), equilibrium is a list \( \langle \mathbf{V}, \mathbf{\tau} \rangle \) satisfying (8)-(10). Taking \( \tau_1 = 1 \) for granted for now (which, as shown below, is valid if \( \rho \) is not too big), let us check the BR condition for \( \tau_0 = 1 \). This reduces to

\[
 r \leq \hat{r}_M \equiv \alpha \sigma (1 - A) (u - c)/c. \tag{11}
\]

If \( r \leq \hat{r}_M \) monetary equilibrium exists, and welfare is higher than barter but lower than credit. Importantly, notice \( \hat{r}_C > \hat{r}_M \) iff \( \mu > 1 - A \), illustrating a result in Kocherlakota (1998): if the monitoring and record-keeping technology — what he calls memory — is perfect in the sense of \( \mu = 1 \), money cannot be essential. This is because \( \mu = 1 \) implies \( \hat{r}_C > \hat{r}_M \), so if money exchange is viable, credit is, too, and the latter is better. Credit is better because with money: (i) potential sellers may have \( a = 1 \); and (ii) potential buyers may have \( a = 0 \).

![Figure 2: Equilibria with Assets as Money](image)

To summarize, fiat money is inessential when \( \mu = 1 \), and essential when \( \mu < 1 - A \), because then for some parameters money works while credit does not. To reiterate, necessary ingredients for essentiality are: (i) not all gains from trade are exhausted by barter; (ii) lack of commitment; and (iii) imperfect monitoring. Note however that while it can be a useful institution, fiat money is in a sense also *tenuous*:

\[\text{While (i) depends on } a \in \{0, 1\}, \text{ with random matching (ii) is robust. With directed matching money can be as good as credit but not better (Corbae et al. 2002, 2003). An exception is when there is value to privacy that makes record keeping undesirable due to, e.g., identity theft (Kahn et al. 2005; Kahn and Roberds 2008). Also, note } \mu \text{ does not appear in (11) because monetary exchange requires no record keeping. Now, one can argue that some record keeping is inherent in monetary exchange, as the fact that someone has currency currently suggests they produced in the past. The counter is that sellers do not care if buyers got assets by fair means or foul – e.g., theft – in the past, and only care about others accepting them in the future.}\]
there is always an equilibrium where it is not valued, plus sunspot equilibria where \( \tau_0 \) fluctuates (Wright 1994). Yet in another sense money is robust: equilibria with \( \tau_0 = 1 \) exist for \( \rho < 0 \) as long as \( |\rho| \) is not too big. To see this, and to check the BR condition for \( \tau_1 = 1 \), Figure 2 shows the equilibrium \( \tau = (\tau_0, \tau_1) \) for any \( \rho \): equilibrium \( \tau = (1, 1) \) exists iff \( |\rho| \) is not too big; \( \tau = (0, 1) \) exists if \( \rho < \rho_c \), in which case buyers are willing to trade assets for goods but sellers will not oblige; and \( \tau = (1, 0) \) exists if \( \rho > \rho_u \), in which case sellers are willing to trade but buyers will not. As shown, these equilibria often coexist, and when the do there is also a mixed-strategy equilibrium.\(^{15}\) A general message is this: whether assets circulate is not always pinned down by primitives: for some \( \rho \) there coexist equilibria where it does and where it does not.

The above discussion concerns symmetric meetings in a large population. Jin and Temzelides (2004) have some meetings involving people who know each other and others involving those who do not. Hence, credit works in some meetings but not others. With a finite set of agents, Araujo (2004) shows that even if deviations cannot be communicated to the population at large, sometimes social norms and contagion strategies can be used to enforce credit: if someone fails to produce for you, stop producing for others, who stop ... and eventually this gets back to the original deviant. This cannot dissuade deviations if the population is large, however, consistent with the stylized fact money is used in large or complicated societies but not in small primitive ones. Even with a large population, we need imperfect monitoring, and \( \mu < 1 \) is just one way of modeling this.\(^{16}\)

In Cavalcanti and Wallace (1999a,b), a fraction \( n_m \) of the population, sometimes called bankers, are monitored in all meetings; the rest, called nonbankers, are never

\(^{15}\)Intuitively, for an arbitrary \( \tilde{\tau} \) used by others, your payoffs are \( V_0(\tilde{\tau}) \) and \( V_1(\tilde{\tau}) \), and (10) gives your BR correspondence, say \( \tau = \Upsilon(\tilde{\tau}) \). Equilibrium is a fixed point. This captures nicely the idea that your individually-optimal trading strategy depends on the strategies of others: when a seller accepts \( a \) he is concerned about getting stuck with it; when a buyer gives up \( a \) he is concerned about getting stuck without it; and these both depend on other agents’ strategies. Note that while the mixed-strategy outcomes in this particular model may not be robust (see fn. 18), it is hard to avoid multiple equilibria in these kinds of models, in general.

\(^{16}\)This random monitoring formulation follows Gu et al. (2013a,b) and Araujo et al. (2015). Cavalcanti and Wallace (1999a,b) have some agents monitored and not others. Sanches and Williamson (2010) have some meetings monitored and not others. Kocherlakota and Wallace (1998) and Mills (2008) assume information about deviations is detected with a lag. Amendola and Ferraris (2013) assume information about deviations is sometimes lost. See also Carmona (2015).
monitored. Agents can now issue notes, pieces of paper with names on them, having no coupon \((\rho = 0)\) but potentially having value in exchange. Notes issued by anonymous agents are never accepted for payment – why produce to get a note when you can print your own for free? – but notes issued by monitored agents may be accepted, which is why these agents might be interpreted as banks. This setup can be used to compare an outside money regime (only fiat currency) to an inside money regime (only notes). The outside money regime is similar to our baseline model, except we can now exploit the fact that some agents are monitored.

Let \(W\) be average utility across monitored and unmonitored agents, and, for illustration, suppose monitored agents never hold money (given \(A\) this does not affect \(W\)). Let us try to implement an outcome where monitored agents produce for anyone that likes their output. For simplicity, set \(\delta = 0\) to rule out barter. Then a monitored agent’s flow payoff is \(\alpha\sigma (n_m u - c)\), since he produces for anyone but only consumes in a meeting with probability \(n_m\), since anonymous agents do not produce without quid pro quo. In some applications, monitored agents can always become anonymous, but suppose here they can be punished by autarky, making it easy to compute their IC. For anonymous agents,

\[
\begin{align*}
    rV_0 &= \alpha\sigma n_m u + \alpha\sigma (1 - n_m) (1 - A) (u + V_1 - V_0) \\
    rV_1 &= \alpha\sigma n_m u + \alpha\sigma (1 - n_m) A (-c + V_0 - V_1),
\end{align*}
\]

which modifies (8)-(9) by recognizing that they can consume but do not have to produce when they meet monitored agent, and \(A\) now denotes the asset supply per nonmonitored agent. Generalizing (11), a monitored agent’s IC is

\[
    r \leq \alpha\sigma (1 - n_m) (1 - A) (u - c) / c.
\]

Notice \(W\) is maximized at \(A^* = 1/2\) (as in the baseline model, this maximizes trade volume). In the other regime, with no outside money, monitored agents can issue notes. This allows them to consume when they meet nonmonitored agents. Also, one can specify that with some probability bankers require nonmonitored agents with notes to turn them over to get goods, to adjust the measure of notes in circulation. Again \(A = 1/2\) is optimal. It is easy to check \(W\) is higher with
inside money, because it lets monitored agents trade more often by issuing notes as needed. While this is not too surprising, the virtue of the method in general is that it allows one to concretely discuss the relative merits of different arrangements.\textsuperscript{17}

These models are in some ways crude.\textsuperscript{18} Yet they nicely illustrate how money can be beneficial and clarify the requisite frictions. In applications, Kiyotaki and Wright (1993), Camera et al. (2003) and Shi (1997b) endogenize specialization to show how money enhances efficiency. Burdett et al. (1995) analyze who searches, buyers or sellers, and Li (1994,1995) considers taxing money to correct externalities with endogenous search intensity. Ritter (1995), Green and Weber (1996), Lotz and Rocheteau (2002) and Lotz (2004) consider introducing new currencies, while Matsuyama et al. (1993) and Zhou (1997) consider international currencies. Corbae and Ritter (2004) analyze credit. Amendola (2008) discusses ways in which one might rule out the nonmonetary outcome, making monetary exchange more robust. Araujo and Shevchenko (2006) and Araujo and Camargo (2006,2008) study learning, which is especially interesting, because with bilateral matching information diffuses through the population slowly due to the same search frictions that are part of what make money useful. There are several models introducing private information (see fn. 58). While simple, simple, based on all this work we contend that these models have many feature that undeniably ring true.

4 The Terms of Trade

Second-generation monetary search theory introduced by Shi (1995) and Trejos and Wright (1995) uses divisible goods and let agents negotiate terms of trade. The

\textsuperscript{17}See Mills (2008), Wallace (2010,2013,2014), Wallace and Zhu (2007) and Deviatov and Wallace (2014) for more applications. Related but different studies of banking in this kind of model include Cavalcanti et al. (1999,2005), Cavalcanti (2004), He et al. (2005) and Lester (2009).

\textsuperscript{18}In particular, everything is indivisible, and that does drive some results. Shevchenko and Wright (2004) argue that any equilibrium with partial acceptability, \( \tau_a \in (0,1) \), is an artifact of indivisibility, but then show how adding heterogeneity yields a similar multiplicity and robust partial acceptability. Note also that indivisibilities introduce a complication that we ignore: as in many nonconvex environments, agents may want to use lotteries, producing in exchange for a probability of getting a (Berentsen et al. 2002; Berentsen and Rocheteau 2002; Lotz et al. 2007). Of course, lotteries are not at all the same as mixed strategies. In Figure 2, e.g., if one allows lotteries what happens is this: for large \( \rho \), the sellers gives his goods to the buyer for sure, and with some probably he gets the asset.
model further illustrates implications for multiplicity, efficiency and dynamics. Using
continuous time, which makes dynamics easier, we have the DP equations
\begin{align*}
r V_0 & = \alpha\delta [u (Q) - c (Q)] + \alpha\sigma A\tau_0\tau_1 [V_1 - V_0 - c (q)] + \dot{V}_0 \tag{15}
\end{align*}
\begin{align*}
r V_1 & = \alpha\delta [u (Q) - c (Q)] + \alpha\sigma (1 - A) \tau_0\tau_1 [u (q) + V_0 - V_1] + \rho + \dot{V}_1, \tag{16}
\end{align*}
where \(\dot{V}_0\) and \(\dot{V}_1\) are derivatives wrt \(t\). These are like (8)-(9), with \(u = u (q)\) the utility
from \(q\) units of consumption and \(c = c (q)\) the disutility of production. Assume
\(u (0) = c (0) = 0, u' (q) > 0, c' (q) > 0, u'' (q) < 0\) and \(c'' (q) \geq 0\) \(\forall q > 0\). Also,
\(\exists \bar{q} > 0\) with \(u (\bar{q}) = c (\bar{q})\). There are two quantities to be determined in (15)-(16), \(q\)
in a monetary trade, and \(Q\) in barter. However, because they are independent, we
focus only on the former (or, we can simply set \(\delta = 0\)).

Before discussing equilibrium, let’s first ask what is incentive feasible. In Section
3, with indivisible goods, credit is viable iff \(r \leq \hat{r}_C\) and money iff \(r \leq \hat{r}_M\), with \(\hat{r}_C\)
given by (7) and \(\hat{c}_M\) is similar except \(1 - A\) replaces \(\mu\). The analog is that we can now
support credit where agents produce \(q\) for anyone that likes their output \(\forall q \leq \hat{q}_C\)
and we can support exchange where agents produce \(q\) for fiat money \(\forall q \leq \hat{q}_M\), where
\(c (\hat{q}_C) = \mu\alpha\sigma u (\hat{q}_C) / (r + \mu\alpha\sigma)\) and \(\hat{q}_M\) is similar except \(1 - A\) replaces \(\mu\). Hence,
IC now impinges on the intensive margin (how much agents trade, not whether
they trade). The applicable version of Koehl (1998) is that \(\mu = 1\) implies
\(\hat{q}_C > \hat{q}_M\). So money is not essential if \(\mu = 1\), but is if \(\mu < 1 - A\), as then some \(q\) can
be supported with money and not with credit.

For equilibrium, there are many alternatives for determining \(q\). Consider first
Kalai’s (1977) bargaining solution, which says that when a buyer gives an asset to
a seller for \(q\), the one who entered the meeting with \(a\) gets a share \(\theta_a\) of the total
surplus.\(^\text{19}\) Since the surpluses are \(S_1 (q) = u (q) - \Delta\) and \(S_0 (q) = \Delta - c (q)\), the
Kalai solution is \(S_1 (q) = \theta_1 [u (q) - c (q)]\), or
\[
\Delta = v (q) \equiv \theta_1 c (q) + \theta_0 u (q), \tag{17}
\]
given the IC’s \(S_1 (q) \geq 0\) and \(S_0 (q) \geq 0\) hold, as they must in equilibrium. Setting
\(\tau_0 = \tau_1 = 1\) and subtracting (15)-(16), we get \(\Delta\) as a function of \(\hat{\Delta}\), and then use
\(^{19}\) Kalai bargaining is very tractable and satisfies the axioms of individual rationality and Pareto
efficiency, which are obviously natural in this context. It also has a strategic foundation (Dutta
2012), although that is not as simple as strategic foundations for Nash bargaining.
(17) to get a simple differential equation \( \dot{q} = e(q) \), where

\[
v'(q) e(q) = [\alpha \sigma(\theta_1 - A) + r \theta_1] c(q) - [\alpha \sigma(\theta_1 - A) - r(1 - \theta_1)] u(q) - \rho.
\]

Letting \( V = (V_0, V_1) \), equilibrium is defined by bounded paths for \( (V, q) \) satisfying these conditions, with \( q \in [0, \bar{q}] \), since that is necessary and sufficient for the IC conditions. Characterizing equilibria involves analyzing this dynamical system.\(^{20}\)

Trejos and Wright (2013) characterize the outcomes. Figure 3 shows the case of a relatively high \( \theta_1 \), with subcases depending on \( \rho \). Starting with \( \rho = 0 \) (middle panel), there are two steady states, \( q = 0 \) and a unique \( q^e > 0 \) solving \( e(q^e) = 0 \). There are also dynamic equilibria starting from any \( q_0 \in (0, q^e) \), where \( q \to 0 \), due to self-fulfilling inflationary expectations. For an asset with a moderate dividend (right panel), \( e(q) \) shifts down, leaving a unique steady state \( q^e \in (0, \bar{q}) \) and a unique equilibrium, since any other solution to \( \dot{q} = e(q) \) exits \([0, \bar{q}] \). If we increase the dividend further to \( \rho > \bar{\rho} \) (not shown), \( e(q) \) shifts further, the steady state with trade vanishes, and assets get hoarded. For a moderate storage cost \( \rho \in (\underline{\rho}, 0) \) (left panel), there are two steady states with trade, \( q_H^e \in (0, \bar{q}) \) and \( q_L^e \in (0, q_H^e) \), plus equilibria where \( q \to q_L^e \) due to self-fulfilling expectations. For \( \rho < \rho \) (not shown), there is no equilibrium with trade and agents dispose of \( a \).

Figure 3: Assets as Money with Divisible Goods and High \( \theta_1 \)

Multiple steady states and dynamics arise because the value of liquidity is partly self-fulfilling: if you think others give low \( q_L \) for an asset then you only give a little

\(^{20}\)One can interpret \( \Delta = v(q) \) as a BR condition to highlight the connection to models presented above. Solve the DP equations for \( \Delta(q) \) for an arbitrary \( q \); then taking \( q \) as given, use the bargaining solution \( v(q) = \Delta(q) \) to get \( q = \Upsilon(q) \) in a meeting. Equilibrium is a fixed point. One usually thinks of best responses by *individuals*, and here it is by *pairs*, but that seems a technicality relative to the conceptual merit of connecting \( q = \Upsilon(q) \) to \( \tau = \Upsilon(\tau) \) in Section 3.
to get it; if you think they give high $q_H$ for it then you give more. Shi (1995), Ennis (2001) and Trejos and Wright (2013) also construct sunspot equilibria where $q$ fluctuates randomly, while Coles and Wright (1998) construct continuous-time cycles where $q$ and $\Delta$ revolve around steady state. This formalization of excess volatility in asset values is different from results in ostensibly similar models without liquidity considerations (Diamond and Fudenberg 1989; Boldrin et al. 1993; Mortensen 1999), which require increasing returns. Here dynamics are due to the self-referential nature of liquidity. Also note that to get an asset, sellers incur a cost above the fundamental value, $\rho/r$. In standard usage this is a bubble: “if the reason that the price is high today is only because investors believe that the selling price is high tomorrow – when ‘fundamental’ factors do not seem to justify such a price – then a bubble exists” (Stiglitz 1990). Now one could argue that liquidity services are fundamental (e.g., Tirole 1985), but rather than discuss semantics, we emphasize the economics: liquidity considerations alone can generate deterministic or stochastic fluctuations.

In terms of efficiency, clearly $q = q^*$ is desirable, where $u'(q^*) = c'(q^*)$. Related to Mortensen (1982) and Hosios (1990), with fiat money and a fixed $A$, one can construct $\theta_1^*$ such that $q = q^*$ in equilibrium iff $\theta_1 = \theta_1^*$, where $\theta_1^* \leq 1$ iff $r$ is not too big. Hence, if agents are patient $\theta_1^* \leq 1$ achieves the first best; otherwise $\theta_1 = 1$ achieves the second best. One can also take $\theta_1$ as given and maximize $W$ wrt $A$.

With $q$ exogenous, as in Section 3, the solution is $A^* = 1/2$. With $q$ endogenous, if $q < q^*$ then $A^* < 1/2$ due to an envelope-theorem argument. This captures in a very stylized way the notion that monetary policy should balance liquidity provision and control of the price level. Although it depends here on the upper bound for $a$, it illustrates the robust idea that it is the distribution of liquidity that really matters.

For a diagrammatic depiction of welfare, let $S_1 = u(q) - \Delta$ and $S_0 = \Delta - c(q)$ denote the buyer’s and seller’s surplus. Any trade must satisfy the IC’s, $S_1 \geq 0$ and $S_0 \geq 0$. The relationship between $S_1$ and $S_0$ as $q$ changes, the frontier of the bargaining set, is $S_0 = -c[u^{-1}(S_1 + \Delta)] + \Delta$ in Figure 4. Kalai’s solution selects the point on the frontier intersecting the ray $S_0 = (\theta_0/\theta_1)S_1$. As $\theta_1$ increases, this ray rotates and changes $S_1$, $S_0$ and $q$. Let $S^* = u(q^*) - c(q^*)$. In the left panel of Figure 4, drawn assuming $\Delta \geq c(q^*)$, $q^*$ can be achieved for some $\theta_1^*$ at the tangency.
between the frontier and the 45° line, and $q$ is too low (high) if $\theta_1$ is below (above) $\theta_1^*$. In the right panel, drawn assuming $\Delta < c(q^*)$, output is too low for all $\theta_1$, and the second best obtains at $\theta_1 = 1$.

Many results do not rely on a particular bargaining solution, and various alternatives used in the literature can be nested by letting $v(q)$ be a generic mechanism describing how much value a buyer must transfer to a seller to get $q$, so the terms of trade solve $\Delta = v(q)$. Consider e.g. generalized Nash bargaining with threat points given by continuation values: $\max_q [u(q) - \Delta]^{\theta_1}[\Delta - c(q)]^{\theta_0}$. The FOC can be written $\Delta = v(q)$ with

\[
v(q) = \frac{\theta_1 u'(q)c(q) + \theta_0 c'(q)u(q)}{\theta_1 u'(q) + \theta_0 c'(q)}.
\]

Notice (18) and (17) are the same at $q = q^*$; otherwise, given $u'' < 0$ or $c'' > 0$, they are different except in special cases like $\theta_\alpha = 1$.

Nash bargaining allows us to show how efficiency depends on various forces. First there is bargaining power; to neutralize that set $\theta_1 = 1/2$. Second there is market power coming from market tightness as reflected in the threat points; to

---

21In terms of the literature, Shi (1995) and Trejos and Wright (1995) use symmetric Nash bargaining; Rupert et al. (2001) use generalized Nash; they all consider threat points given by continuation values and by 0, both of which can be easily derived from strategic bargaining in nonstationary settings (Binmore et al. 1986). Coles and Wright (1998) use a strategic solution for nonstationary equilibria. Trejos and Wright (2013) use Kalai bargaining. Others use mechanism design (Wallace and Zhu 2007; Zhu and Wallace 2007), price posting (Curtis and Wright 2004; Julien et al. 2008; Burdett et al. 2014) or auctions (Julien et al. 2008).
neutralize that set \( A = 1/2 \). Then one can show \( q < q^* \), and \( q \to q^* \) as \( r \to 0 \) (Trejos and Wright 1995). The intuition is simple: In frictionless economies, agents work to acquire purchasing power they can turn into immediate consumption, and hence until \( u'(q^*) = c'(q^*) \). In contrast, in monetary economies, they work for assets that provide consumption only in the future, and hence settle for less than \( q^* \). This comes up with divisible assets, too, but it is a point worth making even with \( a \in \{0,1\} \).


5 Intermediation

Intermediaries, or middlemen, are also institutions that facilitate trade. We now present two models built to study these institutions, one by Rubinstein and Wolinsky (1987) and the other by Duffie et al. (2005). In addition to intermediation being an important topic in its own right, a reason to discuss these models is that, once one gets past notation and interpretation, they work much like the one in Section 4, and we want to make connections between different literature.\(^{22}\)

The Rubinstein-Wolinsky model has three types, \( P \), \( M \) and \( C \), for producers, middlemen and consumers. For simplicity, all agents stay in the market forever (in the original setup, \( P \) and \( C \) exit after trade, to be replaced by clones). The measure of type \( i \) is \( n_i \). There is a divisible but nonstorable good \( q \) anyone can produce at cost \( c(q) = q \) and consume for utility \( u(q) = q \) (i.e., there is transferable utility).\(^{22}\)

\(^{22}\)However, it is again possible to skip to Section 6 without loss of continuity.
Unlike Section 4, there are no gains from trade in \( q \), but it can be used as a payment instrument. There are gains from trade in a different good that is indivisible and storable: type \( C \) consumes it for utility \( u > 0 \), while type \( P \) produces it, at cost \( 0 \), to reduce notation. While \( M \) does not produce or consume this good, he may acquire it from \( P \) to trade it to \( C \). It is not possible to store more than one unit, \( a \in \{0, 1\} \), but the total supply is not fixed, as type \( P \) have an endogenous decision to produce. Also, holding returns are specific to \( P \) and \( M \), where \( \rho_P \leq 0 \) and \( \rho_M \leq 0 \), so that \( -\rho_P \) and \( -\rho_M \) can be interpreted as storage costs.

Let \( \alpha_{ij} \) be the rate at which \( i \) meets \( j \) (there is always a population \( n \) consistent with this). In \( PC \) matches, \( P \) gives the indivisible good to \( C \) for \( q_{CP} \). In \( MC \) matches, if \( M \) has \( a = 1 \) he gives it to \( C \) for \( q_{CM} \). In \( MP \) matches, they cannot trade if \( a = 1 \), and may or may not trade if \( a = 0 \), but if they do \( M \) transfers \( q_{MP} \) to \( P \). Let \( m \) be the measure of type \( M \) with \( a = 1 \). The \( q_i \) that \( i \) gives \( j \) splits the surplus, where \( \theta_{ij} \) is the share or bargaining power of \( i \), with \( \theta_{ji} + \theta_{ij} = 1 \), consistent with Nash or Kalai bargaining since \( u(q) = c(q) = q \). Similar to entry by firms in Pissarides (2000), we need to determine the fraction of \( P \) that produce, say \( \varepsilon \), and the fraction of \( M \) that actively trade, say \( \tau \). In stationary, pure-strategy, asymmetric equilibria, a fraction \( \varepsilon \) of type \( P \) and a fraction \( \tau \) of type \( M \) are always active while the rest sit out. In terms of economics, a key feature is that storage costs are sunk when agents meet, implying holdup problems.\(^{23}\)

Suppose to illustrate the method that \( \tau = 0 \), so \( C \) can only trade directly with \( P \). Then \( rV_C = \alpha_{CP} \varepsilon (u - q_{CP}) \), where \( u - q_{CP} \) is \( C \)'s surplus, since the continuation value \( V_C \) cancels with his outside option \( V_C \). Notice \( \varepsilon \) appears because we assume uniform random matching, in the sense that \( C \) can contact \( P \) even if the latter is not participating (imagine contacts occurring by phone, with \( \alpha_{CP} \) the probability per unit time \( C \) and \( P \) are connected, but \( P \) only picks up if he is on the market). Bargaining implies \( u - q_{CP} = \theta_{CP} u \), where \( u \) is total match surplus since for \( P \) the continuation value also cancels with his outside option. Then \( rV_C = \alpha_{CP} \varepsilon \theta_{CP} u \).

\(^{23}\)Rubinstein-Wolinsky have \( \rho_j = 0 \), but there is still a holdup problem as the transfer from \( M \) to \( P \) is sunk when \( M \) meets \( C \). Given this, they discuss a consignment contract whereby \( M \) transfers \( q \) to \( P \) after trading with \( C \). This is an example of trading arrangements responding endogenously to frictions, but it may or may not be feasible, depending on the physical environment.
Similar expressions hold for $\tau \neq 0$, and for $V_P$ and $V_a$, where the latter is $M$’s value function when he has $a \in \{0,1\}$. In fact, from these DP equations, one might observe this model looks a lot like the one in Section 4, where the indivisible good was money, except here $u(q) = c(q) = q$.\textsuperscript{24}

The BR conditions are standard, e.g., $\varepsilon = 1$ if $V_P > 0$. So is the SS condition. Equilibrium satisfies the obvious conditions, and implies payments in direct trade $q_{CP} = \theta_{PC}u$, wholesale trade $q_{MP} = \theta_{PM}\Delta$, and retail trade $q_{CM} = \theta_{MC}u + \theta_{CM}\Delta$. The spread, or markup, is $q_{CM} - q_{MP} = \theta_{MC}u + (\theta_{MC} - \theta_{MP})\Delta$. Equilibrium exists and is generically unique, as shown in Figure 5 in $(-\rho_P, -\rho_M)$ space (remember that $-\rho$ is the storage cost). When $-\rho_P$ is high, $\varepsilon = 0$ and the market closes. When $-\rho_P$ is low and $-\rho_M$ high, $\varepsilon = 1$ and $\tau = 0$, so there is production but not intermediation. When both are low we get intermediation. For some parameters, $P$ enters with probability $\varepsilon \in (0,1)$, with $m$ adjusting endogenously to make $V_P = 0$. Also, note that when $P$ has a poor storage technology, a low rate of finding $C$, or low bargaining power, \emph{intermediation is essential}: the market opens iff middlemen are active.

\textsuperscript{24}Notice the indivisible good acts a lot like money – a storable object $M$ acquires to get $q$ from $C$ – but Rubinstein-Wolinsky actually call $q$ money, as a synonym for transferable utility. See Wright and Wong (2014) for an extended discussion.
Rubinstein and Wolinsky (1987) use \( \theta_{ij} = 1/2 \ \forall ij \) and \( \rho_P = \rho_M = 0 \). In that case, \( P \) is always active, and \( M \) is active iff \( \alpha_{MC} > \alpha_{PC} \), as is socially efficient. More generally, again related to Mortensen (1982) and Hosios (1990), Nosal et al. (2014,2015) show equilibrium is efficient iff the \( \theta \)'s take on particular values. Related work includes Li (1998), Schevchenko (2004) and Masters (2007,2008); see Wright and Wong (2014) for a recent paper with more citations to related papers on middlemen in similar settings.

A model of intermediation in OTC financial markets by Duffie et al. (2005) provides a natural way to study standard measures of liquidity, like bid-ask spreads, execution delays and trade volume.\(^{25}\) There are agents called \( I \) and \( D \), for investors and dealers. There is a fixed supply \( A \) of an indivisible asset, with \( a \in \{0,1\} \) denoting the asset position of \( I \).\(^{26}\) There is potential bilateral trade when \( I \) meets either another \( I \) or \( D \). Also, \( D \) has continuous access to a competitive interdealer market. In contrast to Rubinstein-Wolinsky, gains from trade emerge not due to \ex\ ante different types (producers and consumers) but due to \( I \)'s valuation of the dividend being hit with idiosyncratic shocks. The flow utility for \( I \) with \( a = 1 \) and valuation \( j \) is \( \rho_j \), where \( j \in \{0,1\} \) and \( \rho_1 > \rho_0 \), and they switch as follows: in any state \( j \), shocks implying \( j' = 1 \) and \( j' = 0 \) arrive at Poisson rates \( \omega_1 \) and \( \omega_0 \).

Again there is a divisible good anyone can consume and produce, for \( u(q) = q \) and \( c(q) = q \) (interpretable as transferable utility, although some people say it loosely represents a bank account that agents can use to save and borrow). It is to be determined if \( I \) trades with \( D \), but \( I \) trades with another \( I \) iff one has \( a = 1 \) and \( j = 0 \) while the other has \( a = 0 \) and \( j = 1 \), a double coincidence. Let \( V_{aj} \) be \( I \)'s value function with asset position \( a \) and valuation \( j \), so that \( \Delta_j = V_{1j} - V_{0j} \) is the value to \( I \) of acquiring the asset when he is in state \( j \). When \( I \) trades with \( I \), the

\(^{25}\)Their search-and-bargaining model can be considered complementary to other approaches in finance, including information- or inventory-based models. It is also has the virtue of realism: “Many assets, such as mortgage-backed securities, corporate bonds, government bonds, US federal funds, emerging-market debt, bank loans, swaps and many other derivatives, private equity, and real estate, are traded in ... [over-the-counter] markets. Traders in these markets search for counterparties, incurring opportunity or other costs. When counterparties meet, their bilateral relationship is strategic; prices are set through a bargaining process that reflects each investor’s alternatives to immediate trade.” (Duffie et al. 2007).

\(^{26}\)In the simplest formulation, \( D \) does not hold the asset, although he can in extensions by Weill (2007,2008). See also Gavazza (2011).
total surplus is $S_I = \Delta_1 - \Delta_0$ and the one that gives up the asset gets a transfer $q_I$ (the subscript indicates $I$ trades with $I$). This yields the party that entered with asset position $a$ a share $\theta_a$, with $\theta_0 + \theta_1 = 1$. The individual gains from trade are $q_I - \Delta_0 = \theta_1 S_I$ and $\Delta_1 - q_I = \theta_0 S_I$, and therefore, $q_I = \theta_0 \Delta_0 + \theta_1 \Delta_1$.

The rate at which $I$ meets $I$ is $\alpha_I$, and the probability that $I$ has asset position $a$ and valuation $j$ is $m_{aj}$. The rate at which $I$ meets $D$ is $\alpha_D$. When $I$ meets $D$, if they trade, one can think of the latter as trading in the interdealer market on behalf of the former for a fee. Obviously this is only relevant when $D$ meets $I$ with $a = 0$ and $j = 1$ or meets $I$ with $a = 1$ and $j = 0$, since these are the only type $I$ agents that currently want to trade. If $I$ gets the asset, $D$ gets $q_A$ (for ask); if $I$ gives up an asset, $D$ gets $q_B$ (for bid). Since the cost to $D$ of getting an asset on the interdealer market is $q_D$, when $D$ gives an asset to $I$ in exchange for $q_A$, the bilateral surplus is $S_A = \Delta_1 - q_D$. And when $D$ gets an asset from $I$ the surplus is $S_B = q_D - \Delta_0$. If $\theta_D$ is $D$’s bargaining power when he deals with $I$, then

$$q_A = \theta_D \Delta_1 + (1 - \theta_D) q_D \text{ and } q_B = \theta_D \Delta_0 + (1 - \theta_D) q_D. \quad (19)$$

One can interpret $q_A - q_D$ as the fee $D$ charges when he gets an asset for $I$ in the interdealer market, and similarly for $q_D - q_B$. The round-trip spread, similar to the markup in Rubinstein-Wolinsky, is $s = q_A - q_B = \theta_D (\Delta_1 - \Delta_0) > 0$.

Trading strategies are summarized by $\tau = (\tau_A, \tau_B)$, where $\tau_A$ is the probability when $D$ meets $I$ with $a = 0$ and $j = 1$ they agree to exchange the asset for $q_A$, while $\tau_B$ is the probability when $D$ meets $I$ with $a = 1$ and $j = 0$ they agree to exchange the asset for $q_B$. The BR conditions are again standard, e.g. $\tau_A = 1$ if $\Delta_1 > q_D$, etc. The reason $I$ with $a = 0$ and $j = 1$ might not trade when he meets $D$ is that $q_D \geq \Delta_1$ is possible. Similarly, $I$ with $a = 1$ and $j = 0$ might not trade when he meets $D$ if $q_D \leq \Delta_0$. For market clearing, because the asset is indivisible, it will be typically the case that $q_D \in \{ \Delta_0, \Delta_1 \}$, and the long side of the market will be indifferent to trade. If the measure of $D$ trying to acquire an asset exceeds the measure trying to off-load one, then $q_A = q_D = \Delta_1$ and $\tau_A \in (0,1)$. In the opposite case, $q_B = q_D = \Delta_0$ and $\tau_B \in (0,1)$. Since the measure of $I$ trying to acquire assets is $m_{10}$ and the measure trying to divest themselves of assets is $m_{01}$, the former is on the short side iff $m_{10} < m_{01}$ iff $A < \hat{A} \equiv \omega_1 / (\omega_0 + \omega_1)$.
The SS and DP equations are standard—e.g., the flow payoff for \( I \) with \( a = 1 \) and low valuation is the dividend, plus the expected value of trading with \( I \) or \( D \), plus the capital gain from a preference shock:

\[
r V_{10} = \rho_0 + \alpha_I m_{01} \theta_1 S + \alpha_D \tau B (q_B - \Delta_0) + \omega_1 (V_{11} - V_{10}) .
\]

An equilibrium is a list \( (V, \tau, m) \) satisfying the usual conditions, and it exists uniquely. The terms of trade are easily recovered, as is the bid-ask spread.

This stylized structure, with a core of dealers and a periphery of investors that may trade with each other or with dealers, is a reasonable representation of many OTC markets. The proportion of intermediated trade is \( \alpha_D / (\alpha_D + \alpha_I) \). If \( \alpha_D \) is small, most exchange occurs between investors, as in markets for specialized derivatives or Fed Funds; small \( \alpha_I \) better approximates NASDAQ. The case \( \alpha_D = 0 \) is interesting for making connections between money and finance: this model has gains from trading \( a \) due to heterogeneous valuations, with \( q \) a payment instrument; the one in Section 4 has gains from trading \( q \), with \( a \) the payment instrument. For intermediation theory, \( \alpha_I = 0 \) is nice since the \( V \)'s and \( q \)'s are independent of \( m \).

This makes it easy to see that spreads are decreasing in \( \alpha_D \) and increasing in \( \theta_D \), and as \( r \to 0 \), \( q_A, q_B \) and \( q_D \) go to the same limit, which is \( \rho_0 / r \) if \( A > \hat{A} \) and \( \rho_1 / r \) if \( A < \hat{A} \). There are also implications for trade volume, often associated with liquidity, but these results are sensitive to \( a \in \{0, 1\} \). The time has come to relax this restriction, first in monetary then financial economics. \(^{27}\)

### 6 The Next Generation

Here we generalize \( a \in \{0, 1\} \) to \( a \in A \) for some less restrictive set \( A \), where one has to somehow handle the endogenous distribution of assets across agents, \( F(a) \). One option is to let \( A = \{0, 1, \ldots, \bar{a}\} \), where \( \bar{a} \) may be finite or infinite, and proceed with a combination of analytic and computational methods. \(^{28}\) Molico (2006) instead lets \( A = [0, \infty) \), and studies the case of fiat currency, allowing the supply to evolve

\(^{27}\)Recent related work on intermediation includes Lester et al. (2013), Babus and Hu (2015), Shen et al. (2015) and Uslu (2015).

\(^{28}\)Results for this case include the following: Under certain assumptions, \( F \) is geometric if \( \bar{a} = \infty \) and truncated geometric if \( \bar{a} < \infty \). One can also endogenize \( \bar{a} \), and for \( \bar{a} < \infty \) and fiat money, one can show the optimum quantity is \( A^* = \bar{a} / 2 \), generalizing Section 3 where \( \bar{a} = 1 \).
according to $A_{+1} = (1 + \pi) A$, where subscript indicates next period, and $\pi > 0$ is the rate of monetary expansion. generated by a lump-sum transfer $T$.\textsuperscript{29}

As above, agents can be buyers or sellers depending on who they meet, but now they are not constrained by $a \in \{0, 1\}$. Maintaining the commitment and information assumptions precluding credit, and setting $\delta = 0$ to eliminate barter, we let $q(a_B, a_S)$ be the quantity and $d(a_B, a_S)$ the dollars traded when the buyer has $a_B$ and the seller $a_S$, assuming for simplicity $(a_B, a_S)$ is observed by both. Then

$$V(a) = (1 - 2\alpha \sigma) \beta V_{+1} (a + T) + \alpha \sigma \int \{u[q(a, a_S)] + \beta V_{+1} [a - d(a, a_S) + T] \} dF(a_S) + \alpha \sigma \int \{-c[q(a_B, a)] + \beta V_{+1} [a + d(a_B, a) + T] \} dF(a_B),$$

(20)

where $V(a)$ is the value function of an agent with $a$ assets. The first term on the RHS is the expected value of not trading; the second is the value of buying from a random seller; and the third is the value of selling to a random buyer.

To determine terms of trade, consider generalized Nash bargaining,

$$\max_{q,d} S_B(q, d, a_B, a_S)^\theta S_S(q, d, a_B, a_S)^{1-\theta},$$

(21)

where $S_B(\cdot) = u(q) + \beta V_{+1}(a_B + T - d) - \beta V_{+1}(a_B + T)$ and $S_S(\cdot) = -c(q) + \beta V_{+1}(a_S + d + T) - \beta V_{+1}(a_S + T)$ are the surpluses. The maximization is subject to $d \leq a_B$. One sometimes hears that this looks like a CIA restriction, but in this context it simply a feasibility condition saying that agents cannot hand over more than they have in quid pro quo exchange. That does not make this a CIA model, because the analysis explicitly describes agents trading with each other, and while it happens that barter and credit are ruled out by the environment in this specification, it is not hard to let some barter or credit back in. That is different from tacking on a

\textsuperscript{29}Up until now we have not mentioned recurring monetary injections, which are awkward when $a \in \{0, 1\}$. However, in Li (1994,1995), Cuadras-Morato (1997), Deviatov (2006) and Deviatov and Wallace (2014), lump-sum transfers combined with asset depreciation can proxy for inflation. Here they are more than a proxy: $T > 0$ translates directly into actual inflation.
CIA restriction over and above the usual budget constraint in classical competitive models (more on this below).

There is a law of motion for $F(a)$ with a standard stationarity condition. A stationary equilibrium is a list $\langle V, q, d, F \rangle$ satisfying these conditions. Molico (2006) analyzes the model numerically. He studies the relationship between inflation and dispersion in prices, defined by $p = d/q$, and asks what happens as frictions vanish. He also studies the welfare effects of inflation. Increasing $A$ by giving agents transfers proportional to their current $a$ has no real effect – it is like a change in units. But a lump-sum transfer $T > 0$ compresses the distribution of real balances, because it raises the price level, and when the value of a dollar falls it hurts those with high $a$ more than low $a$. Since those with low $a$ don’t buy very much, and those with high $a$ don’t sell very much, this compression stimulates economic activity by spreading liquidity around. At the same time, inflation detrimentally reduces real balances, and policy must balance these effects.\footnote{Wallace (2014) conjectures that in any economy with this tradeoff there generically exists an incentive-compatible scheme, with transfers non-decreasing in wealth and not necessarily lump-sum, that is inflationary and raises welfare relative to laissez faire.} Chiu and Molico (2010,2011) extend the analysis to let agents sometimes access a centralized market where they can adjust money balances (more on this below); in one version, they must pay a fixed cost for this access, resembling Baumol’s (1952) inventory approach, but using general equilibrium and not just decision theory.

Those papers focus on stationary equilibria. Chiu and Molico (2014) and Jin and Zhu (2014) extend the methods to study dynamic transitions after various types of monetary injections, and show how the redistributitional impact of policy on $F(a)$ can have rather interesting effects on output and prices. In Jin and Zhu’s formulation, for some parameters, output in a match $q(a_B, a_S)$ is decreasing and convex in $a_S$. This means a policy that increases dispersion in real balances increases average $q$. Now, there are other effects, and this is a numerical result about the net effect (explaining why the findings differ from Molico or Chiu-Molico). The important point is that there are cases where a monetary injection increases dispersion in real balances and hence average $q$, and that leads to slow increases in prices after the injection, where again $p = d/q$. The reason is not that prices are sticky – indeed,
and $d$ are determined endogenously by bargaining in every single trade – but the increase in $q$ keeps $p$ from rising quickly during the transition. This implies that nominal rigidities are not needed to capture time-series observations that suggest money shocks affect mainly output in the short run and prices in the longer run.

A different modeling approach when $\mathcal{A} = [0, \infty)$ tries to harness the distribution $F(a)$ somehow. A method due to Shi (1997a) assumes a continuum of households, each with a continuum of members, to get a degenerate distribution across households. The decision-making units are families, whose members search randomly, as above, but at the end of each trading round they return home and share the proceeds. By a law of large numbers, each family starts the next trading round with the same $a$. The approach is discussed at length by Shi (2006).\(^{31}\) Another method due to Menzio et al. (2012) uses directed search and free entry, so that while there is a distribution $F(a)$, the market segments into submarkets in such a way that agents do not need to know $F(a)$. The useful feature is called block recursivity. Unfortunately, analytical tractability is lost if there is money growth implemented with lump sum transfers, since the real value of the transfer is proportional to aggregate real balances, in which case the model needs to be solved numerically. However, the version in Sun (2012) is tractable even with money growth.

Another method due to Lagos and Wright (2005) delivers tractability by combining search-based models like those presented above with frictionless models. One advantage is that this reduces the gap between monetary theory with microfoundations and mainstream macro.\(^{32}\) Another virtue is realism: Much activity in our


\(^{32}\)That the gap was big is clear from Azariadis (1993): “Capturing the transactions motive for holding money balances in a compact and logically appealing manner has turned out to be an enormously complicated task. Logically coherent models such as those proposed by Diamond (1982) and Kiyotaki and Wright (1989) tend to be so removed from neoclassical growth theory as to seriously hinder the job of integrating rigorous monetary theory with the rest of macroeconomics.” Or, as Kiyotaki and Moore (2001) put it, “The matching models are without doubt ingenious and beautiful. But it is quite hard to integrate them with the rest of macroeconomic theory – not least because they jettison the basic tool of our trade, competitive markets.” The Lagos-Wright setup brings some competitive markets back on board, to continue the nautical metaphor, but also maintains frictions from the search-based approach – one doesn’t want to go too far, since, as Helwig (1993) put it, the “main obstacle” in developing a framework for studying monetary systems is “our habit of thinking in terms of frictionless, organized, i.e. Walrasian markets.”
economic lives is fairly centralized (it is easy to trade, prices are taken parametrically, etc.), as might be approximated by classical GE theory, but there is also much that is decentralized (it is not easy to find counterparties, to use credit, etc.) as in search theory. While one might imagine different ways to combine markets with and without frictions, the Lagos-Wright setup divides each period into two subperiods: in the first, agents interact in a decentralized market, or DM; in the second, they interact in a frictionless centralized market, or CM. Sometimes the subperiods are called “day” and “night” as a mnemonic device, but we avoid this, since other people seem to dislike it. More substantively, we maintain the frictions that make money essential.33

We now go through the details of this model, which has become a workhorse in the literature. DM consumption is still $q$, while CM consumption is a different good $x$.34 For now, $x$ is produced one-for-one using labor $\ell$, so the CM real wage is 1. In the DM, agents can be buyers or sellers depending on who they meet. In the former case period utility is $U(x, 1 - \ell) + u(q)$, and in the latter case it is $U(x, 1 - \ell) - c(q)$, where $U(\cdot)$ is monotone and concave while $u(\cdot)$ and $c(\cdot)$ are as in Section 4. To ease the presentation, we begin with $U(x, 1 - \ell) = U(x) - \ell$ and discuss alternatives later.

The DM value function $V(\cdot)$ is like (20) with one change: wherever $\beta V_{t+1}(\cdot)$ appears on the RHS, replace it with $W(\cdot)$, since before going to the next DM agents now visit the CM, where $W(\cdot)$ is the value function. It satisfies

$$W(a) = \max_{x, \ell, \hat{a}} \{U(x) - \ell + \beta V_{t+1}(\hat{a})\} \text{ st } x = \phi(a - \hat{a}) + \rho a + \ell + T,$$

where $a$ and $\hat{a}$ are asset holdings when trading opens and closes, $\phi$ is the price of $a$ in terms of $x$, $\rho$ is a dividend, and $T$ is a transfer of new money. Note that we allow

33 Lagos and Wright (2007) argue the presence of the CM does not imply agents can communicate DM deviations to the population at large, even if one may get that impression from Aliprantis et al. (2006, 2007a). To clarify, in their model, agents can communicate DM behavior when they reconvene in the CM, which is legitimate, but so is the assumption that agents cannot do so—i.e., spatial separation and limited communication are different frictions, as should be clear from the first- and second-generation models. Moreover, even in their setup, Aliprantis et al. (2007b) show how to ensure money is essential. For more on this see Araujo et al. (2012) and Wiseman (2015).

34 It is easy to make $x$ a vector without changing the results. Also, Julien et al. (2015) study the model where $q$ is indivisible, which delivers some interesting results.
$T < 0$, so lump-sum taxes can be used to contract the money supply. This was not possible in Molico (2006), e.g., because an individual’s $a$ may be too low to pay the tax, but now they can use CM labor $\ell$. Also, nothing changes except $\ell$ if, instead of taxes or transfers, government adjusts the money supply by trading CM goods.

There are constraints $x \geq 0$, $\hat{a} \geq 0$ and $\ell \in [0, 1]$ that we ignore for now. Then, eliminating $\ell$, we have

$$W(a) = (\phi + \rho) a + T + \max_x \{U(x) - x\} + \max_\hat{a} \{-\phi\hat{a} + \beta V_{+1}(\hat{a})\}.$$  \hspace{1cm} (22)

Several results are immediate: (i) $x = x^*$ is pinned down by $U'(x^*) = 1$; (ii) $W(a)$ is linear with slope $\phi + \rho$; and (iii) $\hat{a}$ is independent of wealth. By (iii) we get history independence ($\hat{a}$ is orthogonal to $a$), and hence $F(\hat{a})$ is degenerate when there is a unique maximizer $\hat{a}$. In most applications there is a unique such $\hat{a}$. In some, like Galenianos and Kircher (2008) or Dutu et al. (2012), $F$ is nondegenerate for endogenous reasons, but the analysis is still tractable due to history independence. Similarly, with exogenous heterogeneity (see below), $F$ is only degenerate after conditioning on type, but that is enough for tractability.

It is important to know these results actually hold for a larger class of specifications. Instead of quasi-linear utility, we can assume $U(x, 1 - \ell)$ is homogeneous of degree 1, as a special case of Wong (2012), and hence we can use common preferences like $x^\gamma (1 - \ell)^\gamma$ or $[x^\gamma + (1 - \ell)^\gamma]^{1/\gamma}$. Alternatively, as shown in Rocheteau et al. (2008), the main results all go through with any monotone and concave $U(x, 1 - \ell)$ if we assume indivisible labor, $\ell \in \{0, 1\}$, and use employment lotteries as in Hansen (1985) or Rogerson (1988). Any of these assumptions, $U$ quasi-linear or homogeneous of degree 1, or $\ell \in \{0, 1\}$, imply $\hat{a}$ is independent of $a$. There is one caveat: we need interior solutions at least in some periods.

We now move to the DM, characterized by search and bargaining. Here it makes a difference whether $a$ is real or fiat money. One reason is that in a stationary equilibrium the constraint $d \leq a_B$, which again says agents cannot turn over more than they have, binds with fiat money, but it need not bind with real assets. For now, let $\rho = 0$, so that $d = a_B$. Then consider generalized Nash bargaining,

$$\max_q [u(q) - \phi a_B]^{\theta} [\phi a_B - c(q)]^{1-\theta}. \hspace{1cm} (23)$$
The simplicity of (23) is due to $W'(a) = \phi$ being independent of $a_B$ or $a_S$ (which, by the way, means agents do not need to observe each other’s $a$ to know their marginal valuations). Indeed, (23) is the same as Nash bargaining in Section 4, except the value of the monetary payment is $\phi a_B$ instead of $\Delta$. Hence, the outcome is $\phi a_B = v(q)$ instead of $\Delta = v(q)$, but $v(q)$ is still given by (18).

We can now easily accomodate a variety of alternative solution concepts. For Kalai bargaining, simply use $v(q)$ from (17) instead of (18). Aruoba et al. (2007) advocate the use of Kalai in this model, while recognizing that it has some issues (e.g., interpersonal utility comparisons), for the following reasons: (i) it makes buyers’ surplus increasing in $a$; (ii) it does not give an incentive to hide assets; (iii) it makes $V(a)$ concave; and (iv) it is easy. These results are not always true with Nash bargaining, although Lagos and Rocheteau (2008a) provide a fix. In any case, it seems best to be agnostic and allow different bargaining solutions for different applications.

We can also use Walrasian pricing by letting $v(q) = Pq$, where $P$ is the price of DM goods in terms of $x$ that buyers take parametrically, even though $P = c'(q)$ in equilibrium. To motivate this, Rocheteau and Wright (2005) describe DM meetings in terms of large goups, as opposed to bilateral trade – think about the Lucas-Prescott (1974) search model, as opposed to Mortensen-Pissarides (1994). Or, following Silva (2015) (see also Li et al. 2007,2013), we can use monopolistic competition.35

Here we keep the mechanism $v(q)$ general, imposing only monotonicity plus this condition: $\phi a_B \geq v(q^*)$ implies a buyer pays $d = v(q^*)/\phi$ and gets $q^*$; and $\phi a_B < v(q^*)$ implies he pays $d = a_B$ and gets $q = v^{-1}(\phi a_B)$. One can show this holds automatically for standard bargaining solutions, competitive pricing and many other mechanisms, and indeed it can be derived as an outcome, rather than an assumption, given some simple axioms (Gu and Wright 2015).

35One might say that using perfectly or monopolistically competitive markets means giving up to some extent on microfoundations, compared to search-based models that go into more detail with respect to the meeting/trading process, but obviously it can still be prudent to use these standard solution concepts in applications. Indeed, one can make the DM look even more like a standard competitive market by setting $\alpha \sigma = 1$, to avoid search and matching frictinos. However, $\alpha \sigma < 1$ nicely captures the precautionary demand for liquidity, described by Keynes (1936) in terms of “contingencies requiring sudden expenditure and for unforeseen opportunities of advantage.” Also, $\alpha \sigma < 1$ avoids a problem in many models where velocity must be 1 (Lagos and Rocheteau 2005; Telyukova and Visschers 2013).
For any such $v(q)$, the DM value function satisfies

$$
V(a) = W(a) + \alpha \sigma \{ u[q(a)] - \phi a \} + \alpha \sigma \int \{ \phi \tilde{a} - c[q(\tilde{a})] \} dF(\tilde{a}) ,
$$

(24)

where $a$ is money held by an individual while $\tilde{a}$ is held by others, which we allow to be random at this stage, although in equilibrium $\tilde{a} = A$. Then one can derive

$$
V'(a) = \phi + \alpha \sigma \{ u'[q(a)]q'(a) - \phi \} = \phi \left\{ 1 + \alpha \sigma \frac{u'[q(a)]}{v'[q(a)]} - \alpha \sigma \right\} ,
$$

using $q'(a) = \phi / v'(q)$, which follows from $\phi a = v(q)$. We now insert this into the FOC from the previous CM, $\phi_{-1} = \beta V'(a)$, where the $-1$ subscript indicates last period. The result is the Euler equation

$$
\phi_{-1} = \beta \phi \left[ 1 + \alpha \sigma \lambda(q) \right] ,
$$

(25)

where $\lambda(q) \equiv u'(q) / v'(q) - 1$ is the liquidity premium, or equivalently, the Lagrange multiplier on the constraint in the problem $\max \{ u(q) - v(q) \}$ st $v(q) \leq a \phi$.

Although we focus mainly on stationary equilibrium, for the moment let’s proceed more generally. Using $\phi a = v(q)$, $a = A$, and $A = (1 + \pi) A_{-1}$, (25) becomes

$$
(1 + \pi) v(q_{-1}) = \beta v(q) \left[ 1 + \alpha \sigma \lambda(q) \right] .
$$

(26)

Given any path for $\pi$, one can study dynamics as in Section 4. Various authors have shown there can be cyclic, chaotic and stochastic equilibria, again due to the self-referential nature of liquidity (Lagos and Wright 2003; Ferraris and Watanabe 2011; Rocheteau and Wright 2013; Lagos and Zhang 2013; He et al. 2015).

In a stationary equilibrium $z = \phi A$ is constant, so gross inflation $\phi / \phi_{+1} = 1 + \pi$ is pinned down by the rate of monetary expansion – a version of the quantity equation. Then use the Fisher equation $1 + \iota = (1 + \pi)(1 + r)$ to define $\iota$, which will later serve to reduce notation. There are different ways to think about $\iota$. One is to say that it is the nominal interest rate on a bond that is illiquid (i.e., cannot be traded in the DM), just like saying $r$ is the real interest rate on a bond that is illiquid. This is established practice even when there are no such bonds in the model, since we can always price hypothetical assets. Alternatively, a clean interpretation is to say that $\iota$ is the outcome of a thought experiment: ask agents what nominal payoff
they require in the next CM to give up a dollar in this CM. We can do this for \( r \), too, except using a unit of numeraire and not a dollar. Or, one can simply interpret \( \varrho \) as convenient notation for \( (1 + \pi)(1 + r) - 1 \). In any case, it is equivalent here to describe policy by \( \varrho \) instead of \( \pi \), and then, when \( q^{-1} = q \), to write (26) as

\[ \varrho = \alpha \sigma \lambda(q). \]  

(27)

Given \( \varrho \), (27) determines \( q \). Real balances are then given by \( z = v(q) \) and the price level is \( 1/\phi = A/v(q) \) — another version of the quantity equation. Then, to complete the description of equilibrium, in the CM \( U'(x) = 1 \) determines \( x \), and the individual budget equation determines labor as a function of \( a \), say \( \ell(a) \), where in aggregate \( \int \ell(a) = x \). There is a unique such equilibrium if \( \lambda'(q) < 0 \). For many mechanisms \( \lambda'(q) < 0 \) is automatic (e.g., Kalai and Walras, but not generalized Nash). However, even without \( \lambda'(q) < 0 \), the method in Wright (2010) implies there is generically a unique stationary monetary equilibrium, and it has natural properties like \( \varrho = \alpha \sigma \lambda(q) \).

As for efficiency, with generalized Nash bargaining Lagos and Wright (2005) show \( q = q^* \) under two conditions (see also Berentsen et al. 2007). The first is the Friedman rule \( \varrho = 0 \), which eliminates the intertemporal wedge in money demand. The second is \( \theta = 1 \), again related to Mortensen (1982) and Hosios (1990). Heuristically, when agents bring money to the DM they are making an investment in liquidity, and they underinvest if they do not get the appropriate return. With Nash bargaining this means \( \theta = 1 \). If \( \theta < 1 \) is immutable, it would seem desirable to set \( \varrho < 0 \), but that is not feasible — i.e., there is no equilibrium with \( \varrho < 0 \). This is the New Monetarist version of the zero-lower-bound problem now in vogue. While it has nothing to do with nominal rigidities here, it can still be ascribed to a poor pricing mechanism. Kalai bargaining with any \( \theta > 0 \) delivers \( q^* \) at \( \varrho = 0 \). We return to this in Section 8, where the cost of inflation is considered in more detail.

One can also design a mechanism \( v(\cdot) \) that sometimes delivers \( q^* \) even at \( \varrho > 0 \), as in Hu et al. (2009).\(^{36}\) Their implementation approach characterizes the set

\(^{36}\)One reason this is relevant is that \( \varrho = 0 \) requires deflation, \( \pi = \beta - 1 < 0 \), which means taxation, and as Andolfatto (2013) argues, assumptions that make money essential can make taxes hard to collect.
of stationary incentive-feasible allocations obtained from any mechanism that is individually rational and pairwise efficient. A mechanism is a mapping from the asset holdings in a buyer-seller DM meeting, \((a_B, a_S)\), into a trade, \((q, d)\). Due to lack of monitoring or record keeping, the mechanism cannot use individual trading histories, and cannot induce agents to truthfully reveal past defections from proposed play except through money balances. With a constant money supply, one can show that \((q, \phi)\) is incentive feasible if \(q \leq q^*, c(q) \leq \phi A \leq u(q)\) and \(\alpha \sigma [u(q) - c(q)] \geq r\phi A\), where the last condition says the expected DM surplus is worth the cost of accumulating real balances.

It follows immediately that the efficient \(q^*\) is implementable when

\[ r \leq \alpha \sigma [u(q^*) - c(q^*)] / c(q^*). \]  

Notice (28) is identical to (11) from the second-generation model, except for the term \(1 - A\), since now the asset distribution is degenerate. Hence, if agents are patient, \(q^*\) can be implemented without deflation. Also, thanks to quasi-linearity, mechanism design is especially tractable. Thus, Gu and Wright (2015) show that we can restrict attention to a mechanism \(v(q)\) that is very simple (linear except when IC conditions in the DM would otherwise bind). See Wong (2015) for an analysis without quasi-linearity. In other applications, Rocheteau (2012) studies the cost of inflation, Hu and Rocheteau (2013, 2015) and Araujo and Hu (2015) analyze the coexistence of money and credit or other assets, while Chiu and Wong (2015) investigate alternative payment systems.

That’s the basic model. Many assumptions can be relaxed, but quasi-linearity, or one of the other options mentioned above, is needed for history independence. The framework is of course well posed without that, but then it requires numerical methods. By analogy, while heterogeneity and incomplete markets are worth studying computationally in mainstream macro, it is useful to have standard growth theory as a benchmark to analyze existence, uniqueness vs multiplicity, dynamics and efficiency. The framework just presented is our benchmark for monetary economics.

\footnote{The CM is kept the same, as can be justified by the equivalence between competitive and core allocations. More generally, for a discussion of subtleties in this approach, see Rocheteau (2012), and for additional motivation of the mechanism design approach, see Wallace (2010).}
7 Extending the Benchmark

Some versions of our baseline environment have the CM and DM open simultaneously with agents transiting randomly between them (Williamson 2007); others have multiple rounds of DM trade between CM meetings or vice-versa (Berentsen et al. 2005; Telyukova and Wright 2008; Ennis 2009, Jiang an Shao 2014); others use continuous time (Craig and Rocheteau 2008b; Rocheteau and Rodriguez 2013). An extension in Lagos and Rocheteau (2005) and Rocheteau and Wright (2005), on which we spend more time, has two distinct types, called buyers and sellers because in every DM the former want to consume but cannot produce while the latter produce but do not consume. This provides a natural setting to consider price posting and directed search, sometimes called competitive search equilibrium.

One version of this has market makers set up submarkets in the DM to attract buyers and sellers, and charge them entrance fees, although due to free entry into market making the equilibrium fee is 0. In general, buyers and sellers meet in the DM according to a standard matching technology, where $\alpha(n)$ is the probability a buyer meets a seller and $\alpha(n)/n$ is the probability a seller meets a buyer, with $n$ now denoting the seller/buyer ratio. Notice $\sigma = 1$ here, given directed search avoids matching problems (but see Dong 2011). A submarket involves posting $(q, z, n)$ in the CM, describing the next DM by the terms of trade – agents commit to swapping $q$ units of output for $z = \phi a$ real balances if they meet – as well as $n$ which they use to compute the probability of meeting. Market makers design $(q, z, n)$ to maximize buyers’ surplus subject to sellers getting a minimal surplus of $\bar{S}_S$.

Algebra reduces the market maker problem to

$$\max_{q, z, n} \{\alpha(n) [u(q) - z] - \nu z\} \text{ s.t. } \alpha(n) [z - c(q)] = n \bar{S}_S.$$ (29)

38 The market maker story comes from Moen (1997). We can instead let sellers post the terms of trade to attract buyers, or vice-versa. Often these are equivalent, but not always – Faig and Huangfu (2007) provide an example where they are not equivalent precisely because trade is monetary. Other models with money and competitive search include Faig and Jerez (2006), Huangfu (2009), Dong (2011), Dutu et al. (2011), Bethune et al. (2014) and Choi (2015). Also, note that while some directed search theory takes the matching technology as a primitive, this is not always the case – e.g., see Lagos (2000) or Burdett et al. (2001) for models where it is endogenous. Whether this is an issue depends on the application.
Eliminating \( z \) and taking the FOC’s, we get

\[
\alpha(n) \frac{\partial u}{\partial q} = [\alpha(n) + \ell] \frac{\partial u}{\partial q} \tag{30}
\]

\[
\alpha'(n) [u(q) - c(q)] = \bar{S}_S \left[ 1 + \frac{\ell (1 - \epsilon)}{\alpha(n)} \right], \tag{31}
\]

where \( \epsilon = n \alpha'(n)/\alpha(n) \) is the elasticity of matching wrt participation by sellers. In some applications, it is assumed that all agents participate in the DM for free, so that \( n \) is given. Then (30) determines \( q \), while (31) determines the split of the total surplus between buyers and sellers. In other applications it is assumed that sellers have a cost \( \kappa \) to enter the DM. Then \( \bar{S}_S = \kappa \), and (30)-(31) determine \( q \) and \( n \) jointly. In either case, \( \ell = 0 \) implies \( q = q^* \), and given this, with endogenous entry, it implies \( \alpha'(n^*) [u(q^*) - c(q^*)] = \kappa \).

Lagos and Rocheteau (2005) fix \( n \) but endogenize \( \alpha \) through buyers’ search intensity. With random search and bargaining, the time it takes buyers to spend their money increases with \( \pi \), counter to the well-known hot potato effect of inflation (see, e.g., Keynes 1924, p. 51). This is because \( \pi \), as a tax on money, reduces the DM surplus and hence search effort. But with price posting, although \( \pi \) lowers the total surplus, they show it raises buyers’ share at low \( \pi \). Hence, when \( \pi \) increases buyers spend money faster by increasing search effort. Alternatively, in Liu et al. (2011) buyers spend their money faster at higher \( \pi \), even with random search and bargaining, because they (rather than sellers) have a DM entry decision, and higher \( \pi \) increases \( n \). Relatedly, in Nosal (2011) buyers spend money faster at higher \( \pi \) with random search because higher \( \pi \) reduces their reservation trade, endogenizing \( \sigma \), as Kiyotaki and Wright (1991) did in a first-generation model. See also Li (1995), Shi (1997), Jafarey and Masters (2003), Shi and Wang (2006), Faig and Jerez (2007), Ennis (2009) and Hu et al. (2014). These exercises are germane because they concern duration analysis, for which search theory is well suited, and constitute models of velocity based on explicit search, entry or trading decisions.

Rocheteau et al. (2015) analyze a version of the model with a nondegenerate distribution of \( a \) that is still tractable. Suppose the constraint on labor, \( \ell \leq 1 \)

\[\text{One can show } \ell = 0 \text{ achieves efficiency on both the intensive and extensive margins, } (q^*, n^*). \text{ In contrast, Kalai bargaining with DM entry by sellers yields } n^* \text{ iff } 1 - \theta = \alpha'(n^*) n^* / \alpha(n^*), \text{ the Hosios condition saying that sellers’ share should equal the elasticity of matching wrt their participation. Hence, one might say that competitive search delivers the Hosios condition endogenously.} \]
binds. It then takes several periods for a buyer to accumulate desired real balances. There are some equilibria where, in a DM match with a seller, the buyer spends the money he accumulated so far (there can also be equilibria with partial depletion that need to be analyzed numerically). In these equilibrium $F(a)$ is a truncated geometric distribution, and the value and policy functions can be solved in closed form. The model is tractable thanks to ax ante distinct buyer and seller types because, since because sellers hold $a = 0$, there is only one-sided heterogeneity across matches. Still, equilibrium features a nondegenerate distribution of prices, as in Molico (2006). Under some conditions, inflation can raise output and welfare. Moreover, one can use this model to analytically characterize transitional dynamics following a monetary injection, and show that price adjustments are sluggish in the sense that aggregate real balances increase following the money injection.

Rather than permanently distinct types, it is equivalent to have type realized each period if the realization occurs before the CM closes – (27) still holds, with agents conditioning on all the information they have when they choose $\hat{a}$. Related to this, Berentsen et al. (2007) introduce banking. After the CM closes, but before the DM opens, agents realize if they will be buyers or not, at which point banks are open where they can deposit or withdraw money. Or, to make this look more like standard deposit banking, let everyone put money in the bank in the CM and then, when types are revealed, buyers will withdraw while others leave their deposits alone. A bank contract involves an option to convert on demand interest-bearing deposits into cash, or transfer claims on deposits to a third party. This enhances efficiency because nonbuyers can earn interest on balances that would otherwise lay idle, with interest payments covered by borrowers. Hence, banks reallocate liquidity towards those that could use more, similar to Diamond-Dybvig (1983), although here bankers realistically take deposits and make loans in cash rather than goods.

In some applications, bankers can abscond with deposits or borrowers can renege on debt (Berentsen et al. 2007). Others study the role of banks in investment and growth (Chiu and Meh 2011; Chiu et al. 2013). Related work includes Li (2006,2011), He et al. (2008), Becivenga and Camera (2011) and Li and Li (2013). In some of these, money and banking are complements, since a bank is where one
goes to get cash; in others, they are substitutes, since currency and bank liabilities are alternative payment instruments, allowing one to discuss not only currency but checks or debit cards. Ferraris and Watanabe (2008, 2011, 2012) have versions with investment in capital or housing used to secure cash loans from banks; in these models investment can be too high and endogenous dynamics can emerge. Williamson (2012, 2013, 2014) has models where banks design contracts like Diamond-Dybvig, incorporating multiple assets and private information. His banks hold diversified portfolios that allow depositors to share the benefits of long-term investments with less exposure to liquidity shocks. Ennis (2015) extends that model and uses it to discuss some other contemporary policy issues.40

Returning to participation, as noted above, DM entry decisions can be made by sellers or buyers. We can also assume a fixed population that chooses to be one or the other (Faig 2008; Rocheteau and Wright 2009). These decisions can be inefficient, and in this case \( i > 0 \) might raise welfare (e.g., Aruoba et al. 2007). This is important to the extent that there is a misperception that \( i = 0 \) is always optimal in these models; that is false when there are trading frictions and externalities. And we think this is not trite, like saying that inflation is good because it taxes cash goods, like cigarettes, and someone has decided that people smoke too much. That is trite because, one can ask, why not tax smoking directly? While this is delicate, the frictions that make money essential can arguably make DM taxation difficult. Also, importantly, in New Keynesian models deviations from \( \rho = 0 \) can be desirable due to nominal rigidities, but these distortions can alternatively be fully neutralized by fiscal policy (Correia et al. 2008). Keynesian policy prescriptions follow from

---

40 Gu et al. (2013a) propose a related theory of banking based on commitment. Agents have various combinations of attributes affecting the tightness of their IC’s, and hence determining who should be bankers. Bankers accept and invest deposits, and can be essential (see also Araujo and Minetti 2011). Monitoring can also be endogenized as in Diamond (1984) (see Huang 2015). Moreover, as in some of the settings discussed above, here bank liabilities, claims on deposits, serve to facilitate transactions with third parties – i.e., serve as inside money (see also Donaldson et al. 2015 and Sanches 2015). This is a commonly understood role of banking. Consider Selgin (2007): “Genuine banks are distinguished from other kinds of . . . financial intermediaries by the readily transferable or ‘spendable’ nature of their IOUs, which allows those IOUs to serve as a means of exchange, that is, money. Commercial bank money today consists mainly of deposit balances that can be transferred either by means of paper orders known as checks or electronically using plastic ‘debit’ cards.” This is a particularly natural feature to emphasize in New Monetarist models, given the focus on institutions that facilitate exchange.
their maintained hypotheses of sticky prices, just as ours follow from congestion externalities, *in both cases* only if fiscal policy is limited.

We can add heterogenous DM meetings. Consider a random variable $\psi$, such that when a buyer meets a seller the former gets utility $\psi \cdot u(q)$ from the latter’s goods. This illustrates further differences between the models here and ones with exogenous trading patterns, and how a strict lack of double coincidence meetings is not needed for money to be valued. Suppose in every meeting $i$ and $j$ like each other’s output, but it may be asymmetric, $\psi_i \neq \psi_j$. With pure barter, where $i$ gets $q_i$ from $j$ and $j$ gets $q_j$ from $i$, it is easy to verify that efficiency entails $q_i > q_j$ whenever $\psi_i > \psi_j$. But bargaining yields the opposite, because $j$ drives a harder bargain when $i$ really likes his wares (to say it differently, $q_i$ is really expensive in terms of $q_j$ when $j$ does not like $i$’s goods very much). In a monetary economy, $i$ can pay in $d$ as well as $q$, and this improves his terms of trade. Here agents choose to use cash not because barter is impossible, but because it is coarse.41

We can also make the money supply random. Suppose $A = (1 + \pi) A_{-1}$, where $\pi$ is drawn at the start of the DM. The $\pi$ to be implemented later that period generally affects $\phi$, and hence $q$, but agents do not know it when they chose $\hat{a}$. Thus, agents equate the cost and expected benefit of liquidity. It is still optimal to have $\tau = 0$, but this is no longer equivalent to a unique money supply rule: a given $\tau$ is consistent with any stochastic process for $A$ with the same $E[1/(1 + \pi)]$. Lagos (2010a) characterizes the general class of monetary policies consistent with $\tau = 0$. If $d \leq a$ binds in every state, the results look like those in Wilson (1979) or Cole and Kohlerlakota (1998) for deterministic CIA models. See also Nosal and Rocheteau (2011), where buyers and sellers can be asymmetrically informed about inflation. See also Gu et al. (2015), which analyzes the effects of monetary policy (and other) changes announced at $t$ but not implemented until $t' > t$, and shows that the deterministic (perfect-foresight) transition path between $t$ and $t'$ can involve complicated, nonmonotone, or cycal dynamics.

41 This point was made in second-generation models by Engineer and Shi (1998,2001), Berentsen and Rocheteau (2002) and Jafarey and Masters (2003). Our presentation is based on Berentsen and Rocheteau (2003a). Relatedly, in Jacquet and Tan (2011,2012), money is more liquid than assets with state-dependent dividends, since those are valued differently by agents with different hedging needs, while cash is valued uniformly.
Another extension adds capital, as in Aruoba and Wright (1993), Shi (1997a,b), Aruoba et al. (2011) and others. This can raise questions about competition between \( k \) and \( a \), addressed in Section 9, but here we assume \( k \) is not portable (it cannot be taken to the DM), and claims to \( k \) are not recognizable (they can be costlessly forged), so \( k \) cannot be a medium of exchange. The CM production function is \( f(\ell, k) \), and the DM cost function \( c(q, k) \). Then

\[
W(a, k) = \max_{x, \ell, \hat{a}, \hat{k}} \left\{ U(x) - \ell + \beta V_{+1}(\hat{a}, \hat{k}) \right\}
\]

subject to

\[
x + \hat{k} = \phi(a - \hat{a}) + w_\ell(1 - t_\ell)\ell + [1 + (w_k - \delta_k)(1 - t_k)]k + T,
\]

where \( w_\ell \) and \( w_k \) are factor prices, \( t_\ell \) and \( t_k \) are tax rates and \( \delta_k \) is depreciation. Now \((\hat{a}, \hat{k})\) is independent of \((a, k)\) and \( W \) is linear, generalizing benchmark results. The setup nests conventional real business cycle theory: if \( U(x) = \log(x) \) and \( f(\ell, k) \) is Cobb-Douglas, nonmonetary equilibrium here is exactly the same as Hansen (1985).

At the start of the DM, it is randomly determined which agents want to buy or sell. Then buyers search, while sellers sit on their \( k \) and wait for buyers to show up. The DM terms of trade satisfy \( v(q, k) = \phi\hat{a}/w_\ell(1 - t_\ell) \), where \( v(q, k) \) comes from, say, Nash bargaining. It is not hard to derive the Euler equations for \( a \) and \( k \). Combining these with the other equilibrium conditions, we get

\[
(1 + \pi) v(q, k) = \beta c(q_{+1}, k_{+1})[1 + \alpha \sigma \lambda(q_{+1}, k_{+1})]
\]

\[
U''(x) = \beta U''(x_{+1}) \{1 + [f_1(\ell_{+1}, k_{+1}) - \delta_k](1 - t_k)] - K(q_{+1}, k_{+1}) \}
\]

\[
U'(x) = 1/[(1 - t_\ell)f_1(\ell, k_t)]
\]

\[
x + G = f(\ell, k) + (1 - \delta_k)k - k_{+1},
\]

where \( K(\cdot) \equiv \alpha \sigma \beta [c_2(\cdot) - c_1(\cdot)v_2(\cdot)/v_1(\cdot)] \). Given policy and \( k_0 \), equilibrium is a path for \((q, k, x, \ell)\) satisfying (33)-(36).

The term \( K(\cdot) \) in (34) reflects a holdup problem in the demand for capital, parallel to the one in the demand for money (see also Kurmann 2014 and references therein). The first holdup problem is avoided iff \( \theta = 0 \) and the second iff \( \theta = 1 \). For intermediate \( \theta \), there is underinvestment both in money and in capital, which can have important implications for welfare, and for the model’s empirical performance. In Aruoba et al. (2011), with bargaining, \( \pi \) affects investment by only a little at
calibrated parameters, because the holdup problem makes \( k \) low and insensitive to policy. With price taking (or with directed search and price posting), however, holdup problems vanish, and in the calibrated model eliminating 10\% inflation can increase steady state \( k \) by as much as 5\%. This is sizable, although welfare gains are lessened by the necessary transition. In any case, while we cannot go into all the details, this should allay fears that these models are hard to integrate with mainstream macro or growth theory. They are not.\(^{42}\)

As regards labor markets and monetary policy, theory is flexible. If \( c_k(q,k) = 0 \), the model just presented dichotomizes, with \( q \) solving (33), and \((k,x,\ell)\) solving (34)-(36). This means \( \partial q/\partial \pi < 0 \), but \( \partial k/\partial \pi = \partial x/\partial \pi = \partial \ell/\partial \pi = 0 \). The idea of putting \( k \) in \( c(q,k) \) was precisely to break the dichotomy. It leads to \( \partial \ell/\partial \pi < 0 \). An alternative is to interact \( q \) with \( x \) in utility, say \( U(q,x) - \ell \). Rocheteau et al. (2007) and Dong (2011) do this in models with indivisible labor, where \( 1 - \ell \) is genuine unemployment. They still get \( \partial q/\partial \pi < 0 \), but now, \( U_{12} > 0 \Rightarrow \partial x/\partial \pi < 0 \), and given \( \ell \) is used to make \( x \) this implies \( \partial \ell/\partial \pi < 0 \); symmetrically, \( U_{12} < 0 \Rightarrow \partial \ell/\partial \pi > 0 \). So the Phillips curve can slope up or down, depending on \( U_{12} \). Notice this is a fully-exploitable long-run tradeoff. Even without nominal rigidities, this can rationalize traditional Keynesian prescriptions: if \( U_{12} < 0 \), permanently reducing \( 1 - \ell \) by increasing \( \pi \) is feasible. But it is not desirable, as absent other distortions \( \ell = 0 \) is still optimal. This illustrates a key point. One reason it is worth starting with fundamentals like preferences, rather than taking as a structural relationship the Phillips curve and adopting some loss function over \((\pi,\ell)\), is that it allows us to use properties of \( U \) to evaluate the mechanics and the merits of policy. While an exploitable Phillips curve relationship may exist, exploiting it might be a bad idea.

A different approach to inflation and employment is developed by Berentsen et al. (2011), who add a Pissarides (2000) frictional labor market to our benchmark

\(^{42}\)Aruoba (2011) is a dynamic-stochastic model that performs empirically at business-cycle frequencies about as well as reduced-form models with flexible prices, matching many facts but not, e.g., the cyclicity or persistence of inflation. Venkateswarany and Wright (2013) is a model with money, capital and other assets, designed more for lower frequency observations, where it does fairly well. Aruoba et al. (2015) and He et al. (2015) are versions built to study housing, while Aruoba and Schorfheide (2011) is one with sticky prices. Quantitative work incorporating capital is also performed in the model of Shi (1997a) by Shi (1999a,b), Shi and Wang (2006) and Menner (2006), and in the model of Molico (2006) by Molico and Zhang (2005).
model, let’s call it the LM, convening between the CM and DM. Since labor is allocated in the LM, they make CM utility linear in \( x \), as in typical Mortensen-Pissarides models. Then

\[
W_e(a) = \max_{x, \hat{a}} \{ x + \beta L_e(\hat{a}) \} \text{ st } x = \phi(a - \hat{a}) + \epsilon w_1 + (1 - \epsilon)w_0 + T,
\]

where \( L_e(a) \) is the LM value function indexed by employment status \( e \in \{0, 1\} \), \( w_1 \) is employment income and \( w_0 \) is unemployment income. They assume \( w_1 \) is determined by bargaining in the LM (but paid in the CM, with no loss in generality). Generalizing the benchmark results, \( \hat{a} \) is now independent of \( a \) and \( e \). Unemployed agents in the LM find jobs at a rate determined by a matching function taking unemployment and vacancies as inputs, as is standard.

More novel is the assumption that worker-firm pairs generate output, but do not consume what they produce. Instead, all households and a measure \( n \) of firms participate in the DM, where \( n \) is the employment rate and hence also the measure of firms with workers. The DM arrival rate is determined by a matching function taking as inputs the measures of buyers and sellers. This establishes one link between the LM and DM: higher \( n \) allows buyers to trade more frequently, endogenizing the need for liquidity and thus leading to higher \( q \). A second link is that higher \( q \) leads to higher real revenue for firms, as long as they have some DM bargaining power, which leads to more vacancies and hiring. By taxing real balances, inflation lowers \( q \), and this lowers \( n \). In other words, unemployment and inflation are positively related in the model (although that can be altered by changes in assumptions).

Unemployment is shown in Figure 6 from US data, 1955 to 2005, and from the calibration in Berentsen et al. (2011) generated by inputting actual values of \( \iota \), and counterfactually assuming productivity, taxes and demographics are constant. While it cannot match the 60’s, the theory accounts for much of the broad movement in unemployment over a half century with monetary policy as the only driving force.

It is important to emphasize that in filtered data unemployment comoves with both

---

\( \pi \) and \( \iota \), with the latter probably more relevant, since what matters is expected inflation and for that \( \iota \) may be a better proxy than contemporaneous \( \pi \). Berentsen et al. (2011) argue this using scatter plots, following Lucas (1980), but Haug and King (2014) use more advanced time-series methods and deliver similar results. Because this may not be well known, we extract Figure 7 from their findings, showing filtered inflation and unemployment with a phase shift of 13 quarters, as dictated by fit. We agree with their conclusion that, in terms of the Old Monetarist notion of “long and variable lags” between policies and outcomes, the lag between inflation and unemployment may be long but it is in fact not that variable.44

The framework can also explain sticky-price observations, defined as some firms leaving their nominal prices the same when the aggregate price level increases. In Head et al. (2012), the DM uses price posting à la Burdett and Judd (1983).45 Each seller sets \( p \) taking as given the distribution of other sellers’ prices, say \( G(p) \), and the behavior of buyers. The law of one price fails because some buyers see one but

---

44These findings are related to Old Monetarism in several ways. Friedman (1968) says there can be a trade-off between inflation and unemployment in the short run, but in the longer run the latter gravitates to its natural rate. Friedman (1977) says “This natural rate will tend to be attained when expectations are on average realized ... consistent with any absolute level of prices or of price change, provided allowance is made for the effect of price change on the real cost of holding money balances.” This is the effect operative in the model. The data in Friedman (1977) also suggested to him an upward-sloping Phillips curve, if not as strongly as Figure 7.

45See Wang (2011), Liu et al. (2014), Burdett et al. (2014) and Burdett and Menzio (2014) for related work. Earlier, Head and Kumar (2005) and Head et al. (2010) put Burdett-Judd pricing in the Shi (1997) model, but did not discuss nominal rigidities. The idea here is more closely related to Caplin and Spulber (1987) and Eden (1994), but the microfoundations are different.
others see more than one $p$ at a time. As is standard in Burdett-Judd models, this delivers an equilibrium distribution $G(p)$ on a nondegenerate interval $\mathcal{P} = [\underline{p}, \bar{p}]$. Heuristically, $G(p)$ is constructed so that profit is the same for all $p \in \mathcal{P}$, because low-$p$ sellers earn less per unit but make it up on the volume. This is because buyers that see multiple prices choose the lowest $p$, obviously. Given inflation, sellers that stick to $p$ are letting their real prices fall, but their profits do not fall, because volume increases. Of course, in the longer run they must change, because with persistent inflation $p \notin \mathcal{P}$ eventually.

Figure 7: Shifted/Filtered Data, Unemployment and Inflation.

Figure 8: Price Change Distribution: Model and Data
Although $G(p)$ is unique, theory does not pin down which seller sets which $p$. Head et al. (2012) introduce a tie-breaking rule: when indifferent to change, sellers stick to $p$ with some probability. One can choose this probability to match average duration between $p$ changes in, e.g., Klenow and Kryvtsov (2008) data (sellers end up sticking with probability 0.9). Then there is a unique distribution of $p$ changes. With other parameters calibrated to other moments, the results are shown for the model and data in Figure 8. From this it is apparent there is no puzzle in the price change data, since it is just what simple monetary search models predict. There is also no definitive information about Calvo arrival rates or Mankiw menu costs in these data, since the model makes no use of such devices. This finding should influence the way economists interpret price change data: even if prices look sticky, one cannot conclude central banks should follow policy recommendations based on an alleged cost or difficulty of changing. Again, explicitly modeling the exchange process, and in this case asking why prices are sticky, can matter a lot.46

A similar point concerns studies of optimal monetary and fiscal policy by Aruoba and Chugh (2008) and Gomis-Porqueras and Peralta-Alva (2010). Their results overturn conventional wisdom from the reduced-form literature. One manifestation is that MUF models, e.g., imply $\nu = 0$ is optimal even when other taxes are distortionary under the assumption that utility over goods and money is homothetic (Chari et al. 1991). But Aruoba and Chugh’s (2008) analysis shows that making utility homothetic over goods does not imply the value function is homothetic. Again, the conclusion from these findings is that microfoundations matter.47

46 The model is simultaneously consistent with several facts that are challenging for other approaches: realistic duration, large average price changes, many small changes, many negative changes, a decreasing hazard and behavior that varies with inflation. However, the version in Head et al. (2012) cannot match these plus the macro observations on money demand and the micro observations payment methods. Liu et al. (2014) fix this by introducing costly credit.

47 We cannot go into detail on every application, but we mention a few more. Boel and Camera (2006), Camera and Li (2008), Li (2001,2007), Sanches and Williamson (2010) and Berentsen and Waller (2011a) study interactions between money and bonds or credit. Berentsen and Monnet (2008), Berentsen and Waller (2011b) and Andolfatto (2010,2013) discuss policy implementation. Guerrieri and Lorenzoni (2009) discuss liquidity and cycles. Aruoba et al. (2014) study the effect of inflation on the demand for houses; Burdett Aruoba et al. (2014) study the effect on the demand for spouses. Silveira and Wright (2009) study liquidity in the market for ideas, while growth applications include Waller (2011) and Berentsen et al. (2012). Duffy and Puzzello (2013) experiment in the lab in versions where good outcomes can be supported by social norms, without money, yet find subjects tend to favor money, and interpret this as money acting as a coordination device.
8 The Cost of Inflation

We now go into more detail on the effects of monetary expansion. Equilibrium in
the above models depends on policy, \( \iota \) or \( \pi \), as well as the frictions embodied in \( \alpha \sigma \) and the mechanism \( v(\cdot) \), and these ingredients are all important for understanding
inflation. Consider the typical quantitative exercise in reduced-form models, where
one computes the cost of fully-anticipated inflation by asking how much consumption
agents would be willing to give up to reduce \( \pi \) from 10\% to the level consistent
with the Friedman rule. A consensus answer in the literature is around 0.5\% (e.g.,
Cooley and Hansen 1989 or Lucas 2000; see Aruoba et al. 2011 for a longer list of
references). By contrast, in models along the lines of Lagos and Wright (2005), with
Nash bargaining and \( \theta \) calibrated to match retail markups, eliminating 10\% inflation
is worth around 5.0\% of consumption – an order of magnitude higher.

To explain this, as in Craig and Rocheteau (2008a), Figure 9 generalizes the
welfare-triangle analysis of Bailey (1956). With Kalai bargaining, write (27) as

\[
\iota(z) = \alpha \sigma \left\{ \frac{u'[q(z)]}{v'[q(z)]} - 1 \right\},
\]

which can be interpreted as money demand with \( z = \phi a \) denoting real balances. As
\( \iota \to 0 \), \( z \to z_0 = \theta c(q^*) + (1 - \theta)w(q^*) \), and there is an upper bound \( \bar{\iota} = \alpha \sigma \theta / (1 - \theta) \)
above which \( z = 0 \). If \( \theta = 1 \), then \( z_0 = c(q^*) \) and \( \bar{\iota} = \infty \), and the welfare cost of
going from \( \iota = 0 \) to \( \iota_1 > 0 \) is the area under the curve, \( ABC \), because

\[
\int_{z_1}^{z_0} \iota(z)dz = \alpha \sigma \left\{ u[q(z_0)] - c[q(z_0)] \right\} - \alpha \sigma \left\{ u[q(z_1)] - c[q(z_1)] \right\}.
\]

But if \( \theta < 1 \), the cost of inflation no longer coincides with this area, as buyers receive
only a fraction of the increase in surplus coming from \( z \). The true cost of inflation
is the area \( ADC \), and not \( ABC \), because \( ABC \) ignores the surplus of the seller,
related to the holdup problems discussed above.\(^{48}\)

\(^{48}\)Some mechanisms, including Kalai bargaining or Walrasian pricing, deliver the first best at
\( \iota = 0 \). Hence, by the envelope theorem, a small \( \iota > 0 \) has only a second-order effect. Nash
bargaining with \( \theta < 1 \), however, does not deliver the first best at \( \iota = 0 \), so the envelope theorem
does not apply, and even a small \( \iota > 0 \) has a first order impact. This is an additional effect from
Nash, but even with Kalai bargaining, the cost of inflation is greater than it is with Walrasian
pricing because of the holdup problem.
To quantify the effects, a typical strategy is this: Take $U(x) = \log(x)$, $u(q) = \Gamma q^{1-\gamma}/(1-\gamma)$ and $c(q) = x$. Set $\beta$ to match an average real interest rate. Since it is not so easy to identify, set $\alpha\sigma = 1/2$ and later check robustness (it does not matter much over a reasonable range). Then set $\theta$ to match a markup of 30%. $^{49}$ Finally, set the parameters $\Gamma$ and $\gamma$ to match the empirical money demand curve, i.e., the relationship between $\iota$ and real balances scaled by income. A typical fit is shown in Figure 10, drawn using US data on $M1$ (it would be to use $M1S$ from Cynamon et al. 2006 or $M1J$ from Lucas and Nicolini 2013, but the goal here is mainly to discuss the method). This delivers results close to Lagos and Wright’s 5.0%. Note that the calibration does not imply the DM accounts for a large fraction of total output – it is only about 10%. Note also that as we change the frequency from annual to quarterly or monthly, the relevant values of $\alpha\sigma$ and $\beta$ change, but as long as we are below the bound $\alpha\sigma = 1$, the results are approximately the same – a big advantage of search-based models over those where agents spend all their cash each period.

In an extension with private information, Ennis (2008) finds even larger effects, between 6% and 7%. Dong (2010) allows inflation to affect the variety of goods, and

---

$^{49}$ Lagos and Wright (2005) targeted a markup of 10%, but the results are similar, since 10% already makes $\theta$ small enough to matter. Our preferred 30% is based on the Annual Retail Trade Survey discussed in Faig and Jerez (2005) (although Bethune et al. 2014 say the data imply 39%, the difference does not matter that much for our purposes).
hence exchange on the extensive and intensive margins. She gets numbers between 5% and 8% with bargaining. Wang (2011) uses Burdett-Judd pricing to get price dispersion, so that inflation affects both the average and the variability of prices. He gets 3% in a calibration trying to match dispersion, and 7% when he matches a 30% markup. In a model with capital, Aruoba et al. (2011) get around 3% across steady states, although some of that is lost during transitions. Faig and Li (2009) add a signal extraction problem and decompose the cost into components due to anticipated and unanticipated inflation; they find the former is far more important. Aruoba et al. (2014) add home production, which increases the cost of inflation the same way it magnifies the effects of, e.g., taxation in nonmonetary models. Boel and Camera (2009) consider the distributional impact in a model with heterogeneity. Dutu (2006) and Boel and Camera (2011) consider other countries. Faig and Jerez (2006), Rocheteau and Wright (2009) and Dong (2010) use competitive search, and find costs closer to 1% or 1.5%, since this avoids holdup problems. Relatedly, with capital, in Aruoba et al. (2011) switching from bargaining to Walrasian pricing brings the cost down, even though the effect of \( \pi \) on investment is much bigger with Walrasian pricing. Although the choice of mechanisms matters for welfare results, models with bargaining, price taking and posting can all match the money demand data. Rocheteau (2012) also shows a socially efficient mechanism, as in Hu et al. (2009), can match money demand, but it implies the welfare cost of a
low inflation is nil. Bethune et al. (2014) have informed and uninformed agents, as in Lester (2011), and combine directed search with posting and random search with bargaining. They also introduce the option to use costly credit, and show how all of these ingredients matter quantitatively. All of these results underscore the importance of understanding the microfoundations of information and price formation in decentralized markets.

A more radical extension is Aruoba and Schorfheide (2011), who estimate a model integrating New Monetarist and Keynesian features. They compare the importance of sticky price distortions, which imply $\pi = 0$ is optimal, and the effect emphasized here, which implies $\iota = 0$ is optimal. They estimate the model under four scenarios, depending on the DM mechanism and whether they fit short- or long-run money demand. With bargaining and short-run demand, despite large sticky-price distortions, $\iota = 0$ is optimal. The other scenarios even with parameter uncertainty never imply $\pi = 0$. Craig and Rocheteau (2008b) reach similar conclusions in a menu-cost version of our benchmark model, as in Benabou (1988,1992) and Diamond (1993), except in Diamond-Benabou money is merely a unit of account. It matters: Diamond (1993) argues inflation usefully erodes the market power of sellers; but Craig and Rocheteau show that that is dwarfed by the inefficiency emphasized here for reasonable parameters, making deflation, not inflation, optimal.

However, as mentioned above, sometimes a little inflation can be good. In Craig and Rocheteau (2008b) or Rocheteau and Wright (2009), with endogenous entry or search-intensity, optimal $\pi$ is around 2%. In Venkateswarany and Wright (2013), since capital taxation makes $k$ too low, and this is partially offset by a version of the Mundell-Tobin effect, optimal $\pi$ is around 3.5% (obviously a second-best result). Bethune et al. (2014) can get $\iota > 0$ optimal because equilibrium tends to have too many sellers catering to the uninformed, and since such sellers use more cash, inflation reduces their number. Molico (2006), with a nondegenerate $F(a)$, has examples of a positive redistributive effect, although in Chiu and Molico’s (2010,2011) calibrated models, this effect reduces the cost of inflation but $\pi < 0$ is still optimal. The same is true of Dressler (2011a,b), although his majority-voting equilibrium has $\iota > 0$. Based on all this, while understanding the effects of inflation is
an ongoing project, progress to date provides little support for the dogmatic position in some Keynesian research that monetary considerations of the type considered here are irrelevant and can be ignored in policy discussions.

9 Liquidity in Finance

Assets other than currency also convey liquidity. To begin this discussion, we emphasize that assets can facilitate transactions in different ways. First, with perfect credit, there is no such role for assets. Perfect credit means default is not an option. But if sellers worry buyers will renege, they may insist on getting something, like an asset, by way of quid pro quo. Or, instead, as in Kiyotaki and Moore (1997), they may require assets to serve as collateral that can be seized to punish default.

To pursue this, first consider our benchmark model with perfect credit. Then

\[ W(a, D) = \max \{ U(x) - \ell + \beta V(\hat{a}) \} \quad \text{st} \quad x = \phi(a - \hat{a}) + \rho a + \ell + T - D, \]

where \( D \) is debt from the last DM, denominated in numeraire (one-period debt is imposed without loss of generality). Any \( q \) can be purchased in the DM if a buyer promises to make a payment in the CM of \( D = v(q) \). For any Pareto efficient mechanism, \( q = q^* \) and \( V(a) = W(a, 0) + \alpha \sigma [u(q^*) - c(q^*)] \). Since \( a \) does not affect DM trade, the Euler equation is \( \phi_{-1} = \beta (\rho + \phi) \), the only bounded solution to which is the constant solution \( \phi = \phi^* \equiv \rho/r \), where \( \phi^* \) is the fundamental price.

Now let debtors default. Also, suppose we cannot take away defaulters’ future credit. Then the only punishment is seizing assets pledged as collateral - assuming,

---

50 Early papers where real assets facilitate trade in our benchmark model include Geromichalos et al. (2007) and Lagos (2010a,2010b,2011), who have equity \( a \) in fixed supply, and Lagos and Rocheteau (2008a), who have capital \( k \). See also Shi (2005), Rocheteau (2011), Lester et al. (2002), Li et al. (2012), Rocheteau and Petrosky-Nadeau (2012), Nosal and Rocheteau (2013), Hu and Rocheteau (2013) and Venkateswarany and Wright (2013). There are applications using such models to study financial issues like the credit-card-debt puzzle (Telyukova and Wright 2008), on-the-run phenomena (Vayanos and Weill 2008), the equity-premium and risk-free-rate puzzles (Lagos 2008,2010b), home bias (Geromichalos and Simonovska 2014), repos (Narajabad and Monnet 2012), the term structure (Geromichalos et al. 2013; Williamson 2013) and housing bubbles (He et al. 2015).

51 In the notation of Sections 3 and 4, \( \mu = 0 \). Perhaps defaulters can move to new towns, e.g., where they are anonymous. However, this is not meant to diminish the importance of taking away defaulters’ future credit, in general. New Monetarist models of credit based on Kehoe and Levine (1993) instead of Kiyotaki and Moore (1997) include Gu et al. (2013a,b), Bethune et al. (2015) and Carpella and Williamson (2015).
say, that they have been assigned to third parties with commitment. As in standard
Kiyotaki-Moore models, borrowers can pledge only a fraction $\chi \leq 1$ of their assets,
with $\chi$ exogenous (it is endogenized in Section 10). The IC for honoring obligations
is $D \leq (\phi + \rho) \chi a$, as for higher $D$ it is better to forfeit the collateral. Hence, the
Kiyotaki-Moore debt limit is $\bar{D} = (\phi + \rho) \chi a$. But as noted in Lagos (2011), rather
than using the assets to secure his promise, a buyer can hand them over and finalize
the transaction in the DM. It is equivalent for assets to serve as a medium of exchange
or as collateral. One can imagine exceptions — e.g., if it is “inconvenient” to use
part of your house as a payment instrument, you may prefer to get a home-equity
loan — but unless that is modeled explicitly, secured credit à la Kiyotaki-Moore is
not distinct from assets serving as media of exchange à la Kiyotaki-Wright.\footnote{One can also interpret it as a repo: the buyer gives up assets for $q$ in the DM, then buys
them back with numeraire in the CM. The point is not that repos are essential, only that there
are different ways to decentralize trade, or at least different ways to describe it.}

An ostensible distinction is that with secured credit you can only pledge a fraction
$\chi$ of assets, but we can just as well say you can only hand over a fraction $\chi$ of your
assets, and again we show how this can arise endogenously in Section 10. While one
might tell different stories, the equations are the same. Another such distinction is
that the Kiyotaki-Moore literature usually talks about producer credit, not consumer
credit, but in terms of theory, as we said earlier, that is merely a relabeling. A less
delusory distinction may be this: In the models presented above, assets are useful
for the acquisition of $q$. Suppose what you want is not $q$, but more of the same
asset, like a producer increasing business capital or a homeowner increasing housing
capital. It will not do to exchange old $k$ for new $k$. But it might be useful to pledge
the old $k$ to get new $k$ on credit if we assume the former is pledgeable but the latter
is not, which is perhaps arbitrary but not crazy.

In any event, let us include the parameter $\chi$, and consider assets with $\rho > 0$.
Now we cannot be sure $d \leq \chi (\phi + \rho) a$ binds. If it binds, the analog of (25) is

$$\phi_{-1} = \beta (\phi + \rho) [1 + \alpha \sigma \chi \lambda (q)] ,$$

(37)

where $\lambda (q)$ is as before. Using $v (q) = \chi (\phi + \rho) a$ to eliminate $\phi$ and $\phi_{-1}$, we get a
difference equation in $q$ analogous to (26). While $\rho > 0$ rules out equilibria where
Figure 11: Asset demand and supply

\( \phi = 0 \) or \( \phi \to 0 \), there can still exist cyclic, chaotic and stochastic equilibria (Lagos and Wright 2003; Zhou 2003; Rocheteau and Wright 2013). Thus, it is not the fiat nature of money that generates dynamics, but an inherent feature of liquidity, which applies to assets serving as a means of payment or collateral, whether they are nominal or real. We claim that \( \phi \) is nonmonotone in \( \chi \): first, \( \phi = \phi^* \) when \( \chi = 0 \) and when \( \chi \) is so big that the liquidity constraint is slack; then, since \( \phi > \phi^* \) for intermediate values of \( \chi \), the claim follows. Hence, as assets become more pledgeable, there can emerge an endogenous price boom and bust (He et al. 2015).

To check if the liquidity constraint binds, assume that it does, and use \( v(q) = \chi(\phi + \rho)a \) to eliminate \( q \) from (37). In stationary equilibrium, the result can be interpreted as the long-run demand for \( a \) as a function of \( \phi \). One can show demand is decreasing for \( q < q_0 \), defined by \( \lambda(q_0) = 0 \), even if \( \lambda(q) \) is not monotone (Wright 2010). Then define \( A^* \) by \( v(q_0) = \chi(\phi^* + \rho)A^* \), where \( \phi^* = \rho/r \), so that \( a = A^* \) satiates a buyer in liquidity. The resulting asset demand curve is shown in Figure 11, which truncates (37) at \( \phi^* \). It is now immediate that \( A \geq A^* \) implies liquidity is plentiful and \( \phi = \phi^* \), while \( A < A^* \) implies liquidity is scarce and bears a premium \( \phi > \phi^* \). We also mention that some applications use one-period assets instead of long-lived assets. Then (37) changes to \( \phi_{-1} = \beta \rho [1 + \alpha \sigma \chi \lambda(q)] \) where now \( v(q) = \chi \rho a \), since the asset pays a dividend \( \rho \) but has no resale value in the next CM; otherwise the results are similar.
The above analysis applies to assets in fixed supply. Neoclassical capital is similar except the supply curve is horizontal instead of vertical, so liquidity considerations are manifest not by $\phi > \phi^*$ but by $k > k^*$. Housing is similar when it conveys liquidity via home-equity lending, except supply need not be horizontal or vertical. Recent research (see fn. 50) studies models with different combinations of these kinds of assets plus currency. One application concerns open market operations, or OMO’s, in economies with money plus a real, one-period, pure-discount government bond issued in the CM and paying a unit of numeraire in the next CM. The following discussion encompasses features in Williamson (2012, 2013, 2014), Rocheteau and Rodriguez (2013) and Rocheteau et al. (2014).

Suppose these bonds are partially liquid – i.e., can be traded in some DM meetings. Bonds affect the government budget equation, but as in the benchmark model, $T$ adjusts to satisfy this after a policy change. The nominal returns on illiquid real and nominal bonds are still $1 + r = 1/\beta$ and $1 + \tau = (1 + \pi)/\beta$. The nominal return on the liquid real bond – i.e., the amount of cash accruing in the next CM from a dollar put into these bonds in the current CM – is denoted $\nu_b$, and generally differs from $\nu$. Suppose in the DM, with probability $\alpha_m$ a buyer meets a seller that accepts only $a_m$, and they trade $q_m$; with probability $\alpha_b$ he meets one that accepts only $a_b$, they trade $q_b$; and with probability $\alpha_2$ he meets one that accepts both, they trade $q_2$. One can also say that conditional on being accepted, $\chi_m$ and $\chi_b$ are the fractions of the buyer’s $A_m$ and $A_b$ that can be used in the transaction. While $\chi_m = \chi_b = 1$ is a fine special case, again, these are endogenized in Section 10, and it is shown how $\chi_j < 1$ can emerge in private information settings where sellers are worried about counterfeits.

In any case, buyers always spend all the cash they can, but may or may not use all their bonds. Suppose they do use all their bonds in type-$b$ and type-2 meetings.

---

53See also Dong and Xiao (2014), Han (2014) and Ennis (2015). Also, while these papers consider both real and nominal bonds, we concentrate here solely on the former, but most of the results are similar.

54While it may be unusual for households to use bonds as payment instruments, it is common for financial firms to use them as collateral, and just like we can reinterpret agents as producers instead of consumers, many applications interpret them as banks. See Berentsen and Monnet (2008), Koeppel et al. (2008, 2012), Martin and Monnet (2010), Chapman et al. (2011, 2013), Berentsen et al. (2014), Bech and Monnet (2015) and Chiu and Monnet (2015). See also Section 11.
which happens when liquidity is scarce. The Euler equations are

\[ t = \alpha_m \chi_m \lambda(q_m) + \alpha_2 \chi_m \lambda(q_2) \]

\[ s = \alpha_b \chi_b \lambda(q_b) + \alpha_2 \chi_b \lambda(q_2), \]

where \( s \equiv (t - \tau_b) / (1 + \tau_b) \) is the spread between illiquid and liquid bond returns.\(^{55}\) As Rocheteau et al. (2014) point out, \( \tau_b < 0 \) is possible in this model, although we still have \( t \geq 0 \) for illiquid bonds. When \( \tau_b < 0 \), agents are willing to hold \( a_b \) because its expected liquidity benefit exceeds that on cash, or because it is more pledgeable, and note that this does not violate no-arbitrage as long as bonds issued by private agents in the CM do circulate in the DM. Whether or not it is empirically relevant, this shows how \( \tau_b < 0 \) is a logical possibility (see also He et al. 2008, Sanches and Williamson 2010 or Lagos 2013a).

The effects of \( \pi \) or \( \tau \) are still equivalent since, again, \( 1 + \tau = (1 + \pi) / \beta \). As long as \( \alpha_2 > 0 \), higher \( \tau \) raises \( \phi_b \) and \( s \) as agents try to shift out of cash and into bonds. The effect of \( \tau \) on \( \tau_b \) is nonmonotone: a Fisher effect raises the nominal return for a given real return, but a Mundell effect lowers the real return as demand for bonds rises, and either can dominate. For OMO’s, note that buying bonds with cash is formally equivalent to buying them with general tax revenue: higher \( A_m \) means higher prices but the same \( \phi_m A_m \). Hence, any real impact comes from changing \( A_b \). One can show \( \partial z / \partial A_b < 0 \) and \( \partial q_m / \partial A_b < 0 \) if \( \alpha_2 > 0 \). Intuitively, higher \( A_b \) makes liquidity less scarce in type-2 meetings, so agents try to economize on cash, but this comes back to haunt them in type-\( m \) meetings. Since an OMO that raises \( A_m \) results in higher \( z \), prices go up by less than \( A_m \). One could misinterpret this as sluggish prices. In fact, \( z \) increases because \( A_b \) falls. Similarly, this OMO leads to a reduction in \( \tau_b \). One could misinterpret this by thinking the increase in \( A_m \) reduces nominal returns, but again lower \( \tau_b \) actually happens because \( A_b \) falls. Hence, it is not so easy to make inferences about monetary neutrality. For some

\(^{55}\) This is related to what Krishnamurthy and Vissing-Jorgensen (2012) call the “convenience yield” and measure by the difference between government and corporate bond yields (see also Nagel 2014). One might say the model here rationalizes their reduced-form specification with T-bills in the utility function, but it’s not clear if that’s desirable, any more than saying the models presented above rationalize money in the utility function. The sine qua non of our approach is modeling exchange explicitly, not deriving indirect utility functions from primitives.
parameterizations, as Williamson (2012) emphasizes, the model can also generate a liquidity trap, with low $q$'s, and ineffectual OMO's, because changing $A_b$ crowds out real balances, with total liquidity staying put, as in Wallace (1981,1983).

At this juncture we offer some general comments on the coexistence of money and riskless, perfectly-recognizable bonds with positive returns. Papers trying to address this coexistence (or rate of return dominance) puzzle generally impose some asymmetry between money and bonds. Aiyagari et al. (1996), e.g., study a second-generation model with money and two-period government bonds. As in Li (1994,1995), Aiyagari and Wallace (1997) or Li and Wright (1998), there are government agents that act like anyone else except: in meetings with private sellers, they may either pay with cash or issue a bond; they may refuse to accept not-yet-matured bonds from private buyers; and in any meeting they can turn matured bonds into cash. Equilibria with valued money and interest-bearing bonds exist because of asymmetry in the way government treats the assets.\footnote{A different asymmetry is studied in Zhu and Wallace (2007) and Nosal and Rocheteau (2013), where the trading protocol gives a larger share of the surplus to agents with more money.}

Such asymmetries are adopted because of a belief that, under laissez faire, absent exogenous assumptions that favor money there are no equilibria where it coexists with default-free, interest-paying, nominal bearer bonds. Yet arguably there are episodes where such securities and money both circulated, a strong instance of the rate-of-return-dominance puzzle. Lagos (2013a) addresses this in a version of the benchmark model where currency consists of notes that are heterogeneous in extraneous attributes – e.g., serial numbers – called moneyspots, to make a connection with sunspots. These payoff-irrelevant characteristic are enough to get money coexisting with interest-bearing bonds. Heuristically, the extraneous attributes are priced, so that different notes are valued differently, supported purely by beliefs (see Lagos 2013a,b and Wallace 2013 for more discussion).

Moving from pure theory to more applied issues, Lagos (2008,2010b) shows how liquidity helps us understand two of the best-known issues in finance, the risk-free-rate and the equity-premium puzzles. There are two real assets, a one-period risk-free government bond, and shares in a tree with random dividend $\rho$, with prices $\phi_b$ and $\phi_s$. In a minor modification of the benchmark model, returns are in terms
of a second CM consumption good $y$, which is numeraire, while $p_x$ is the price of $x$, and CM utility is $U(x, y)$. For now both assets can be used in all DM transactions, $\alpha_2 > 0 = \alpha_b = \alpha_s$ (but see below). Since feasibility implies $y = \rho$, letting $x(y)$ solve $U_1(x, y) = 1$, we have the Euler equations

$$U_2 \left[ x(\rho_{-1}), \rho_{-1} \right] \phi_{b, -1} = \beta \mathbb{E} \{ U_2 [x(\rho), \rho] [1 + \alpha_2 \lambda(q)] \} \quad (40)$$

$$U_2 \left[ x(\rho_{-1}), \rho_{-1} \right] \phi_{s, -1} = \beta \mathbb{E} \{ U_2 [x(\rho), \rho] (\phi_s + \rho) [1 + \alpha_2 \lambda(q)] \} \quad (41)$$

Compared to our benchmark, now expected marginal utility $U_2(\cdot)$ at different dates matters; compared to asset-pricing models following Mehra and Prescott (1985), now liquidity matters. From (40)-(41) follow a pair of restrictions,

$$\mathbb{E}(\Omega R_b - 1) = \xi_b \quad (42)$$

$$\mathbb{E}[\Omega(R_s - R_b)] = \xi_s, \quad (43)$$

involving the MRS, $\Omega \equiv \beta U_2(x, \rho)/U_2(x_{-1}, \rho_{-1})$, and measured returns, $R_b \equiv 1/\phi_{b, -1}$ and $R_s \equiv (\phi_s + \rho)/\phi_{s, -1}$. For Mehra-Prescott, $\xi_b = \xi_s = 0$, and for standard preferences this is violated by data, where $\xi_b < 0$ and $\xi_s > 0$. By contrast, here

$$\xi_b = -\alpha_2 \mathbb{E} [\Omega \lambda(q) R_b] \quad (44)$$

$$\xi_s = \alpha_2 \mathbb{E} [\Omega \lambda(q) (R_b - R_s)]. \quad (45)$$

From (44), $\xi_b < 0$ as long as $\alpha_2 \lambda(q_{t-1}) > 0$, which says the bond has liquidity value in some state of the world. Hence, a liquidity-based model is always at least qualitatively consistent with the risk-free-rate puzzle. Also, $\xi_s$ is the weighted average return differential $R_b - R_s$ across states. This is generally ambiguous in sign, but suppose there is a high- and a low-growth state, and $R_b - R_s$ is positive in the latter and negative in the former. If the weight $\Omega \lambda(q)$ tends to be larger in the low state, then $\xi_s > 0$. Hence, even without giving bonds a liquidity advantage over equity, the model rationalizes the equity-premium puzzle. In a calibration with $U(x, y) = \hat{U}(x) + y^{1-\gamma}/(1 - \gamma)$, the model is quantitatively consistent with both puzzles for $\gamma = 10$, while standard calibrations of Mehra-Prescott need $\gamma = 20$. Modest differences in acceptability matter a lot. Suppose buyers can only use bonds in 2% of DM trades, and can use bonds or shares in 98%. Then $\gamma = 4$ generates an equity premium of 7%, which is 10 times larger than Mehra-Prescott with $\gamma = 4$. 59
These models have many other implications for financial economics and policy analysis. In Lagos (2010a, 2011), e.g., currency and claims to a random aggregate endowment can both be used in the DM, and in some states, asset values are too low to get $q^*$. Then monetary policy can mitigate this by offsetting liquidity shortages. More generally, the message is that liquidity considerations have important implications for the effects of monetary policy on asset markets. This is no surprise; we are simply suggesting a tractable GE framework that makes this precise.57

10 Information and Liquidity

Going back to Law (1705), Jevons (1875) and others, one approach to understanding the moneyness of assets appeals to informational frictions.58 Alchian (1977) iconoclastically goes so far as to say “It is not the absence of a double coincidence of wants, nor of the costs of searching out the market of potential buyers and sellers of various goods, nor of record keeping, but the costliness of information about the attributes of goods available for exchange that induces the use of money.” While all of the models discussed above involve some information frictions, to hinder credit, here we have in mind private information about the quality of goods or assets.

A first-generation model by Williamson and Wright (1994) has no double coincidence problem, but barter is hindered by goods that are lemons – i.e., they are cheaper to produce but provide less utility. Sometimes an agent recognizes quality before accepting a good, and sometimes not (it is match specific). Depending on

57 In related work, Lagos (2010a, 2011) shows a large class of state-contingent policies implement $\alpha = 0$ and make asset prices independent of monetary considerations. But one can instead target a constant $\alpha > 0$, implying asset prices depend on policy and can persistently deviate from fundamental values. To support a constant $\alpha$ the growth rate of $A_m$ is low in states where real asset values are low, introducing a negative relationship between $\alpha$ and returns. Even if variations in output are exogenous, a positive correlation between inflation and output emerges.

parameters, it is often the case that in equilibrium agents accept goods they do not recognize: Suppose otherwise; then since agents who do recognize low-quality goods never accept them, agents with lemons cannot trade; therefore no one produces lemons and hence you can accept goods with impunity even if you cannot evaluate their quality before trading. In this situation, equilibrium entails mixed strategies, where sellers produce low quality, and buyers accept unrecognized goods, with positive probabilities. Fiat currency can improve welfare, because in monetary equilibria the incentive to produce high quality can be higher.

Moving to private information about asset quality, consider our benchmark economy with one-period-lived assets in fixed supply $A$, with payoff $\rho \in \{\rho_L, \rho_H\}$, where $\Pr(\rho = \rho_H) = \zeta$ and $\Pr(\rho = \rho_L) = 1 - \zeta$. Assume $\rho$ is common to all units of the asset held by an agent, so it cannot be diversified, but it is independent across agents — e.g., the payoff depends on local conditions specific to the holder. The asset holder has private information about $\rho$ as in Plantin (2009). Suppose the holder in a meeting makes a take-it-or-leave-it offer. Using Cho and Kreps’ (1987) refinement, Rocheteau (2011) shows the equilibrium is separating: holders of $L$-type assets make the full-information offer, $c(q_L) = \min\{\rho_L a, c(q^*)\}$ and $d_L = c(q_L)/\rho_L$; and holders of $H$-type assets make an offer satisfying

$$\rho_H d_H \geq c(q_H)$$

$$u(q_H) - \rho_L d_H \leq u(q_L) - c(q_L).$$

In particular, (47) says $L$-type buyers have no incentive to offer $(q_H, d_H)$.

The least-cost separating offer satisfies (46)-(47) at equality, so $d_H = c(q_H)/\rho_H \in (0, d_L)$, while $q_H \in [0, q_L)$ solves $u(q_H) - c(q_H)\rho_L/\rho_H = u(q_L) - c(q_L)$. Thus, $H$-type buyers retain a fraction of their holdings as a way to signal quality, and hence $q_H < q_L$ (interpretable as over-collateralization, as in DeMarzo and Duffie 1999). When the $L$-type asset is a pure lemon, $\rho_L = 0$, both $q_L$ and $q_H$ tend to 0.59 In contrast to models with exogenous constraints, agents here turn over all their assets in trade with probability $1 - \zeta$, and a fraction $d_H/a$ with probability $\zeta$, where $d_H/a$ depends

---

59 This is the case in Nosal and Wallace (2007), Shao (2009) and Hu (2013). The equilibrium, however, would be defeated in the sense of Mailath et al. (1993). A perhaps more reasonable outcome is the best pooling equilibrium from the viewpoint of a buyer with the $H$-type asset, where $q$ solves $\max\{u(q) - \rho_H d\}$ st $\zeta \rho_H d = c(q)$; this still implies $q$ is inefficiently low.
on $\rho_L$ and $\rho_H$. With a one-period-lived asset, while $\phi^* = \beta [\zeta \rho_H + (1 - \zeta) \rho_L]$ is the fundamental value, here we get

$$\phi = \phi^* + \beta \alpha \sigma \theta \left\{ \frac{\zeta \rho_H \lambda(q_H) \rho_L \lambda(q_L)}{\rho_H \lambda(q_H) + \rho_H - \rho_L} + (1 - \zeta) \rho_L \lambda(q_L) \right\}. $$

The liquidity premium $\phi - \phi^*$ depends on $\rho_H - \rho_L$. As $\rho_L \rightarrow \rho_H$, the premium goes to $\beta \alpha \sigma \theta \lambda(q)$. As $\rho_L \rightarrow 0$, the asset becomes illiquid and $\phi \rightarrow \phi^*$. If the asset is abundant, $A \rho_L \geq c(q^*)$, the liquidity premium is 0 even though $q_H < q_L = q^*$. Thus, $q_H$ can be inefficiently low even with abundant assets when they are imperfectly recognizable, meaning other assets may also play an essential role. In Rocheteau (2008), the second asset is fiat money. If it is perfectly recognizable then the same logic applies: agents with $L$-type assets make the complete-information offer while those with $H$-type assets make an offer that others do not want to imitate. When $\iota \rightarrow 0$, agents hold enough currency to buy $q^*$, and no one uses the asset in DM trade. For $\iota > 0$ they do not hold enough currency to buy $q^*$, and spend it all plus a fraction of their risky assets. Asset liquidity as measured by this fraction clearly depends on monetary policy.\textsuperscript{60}

Asset quality can also depend on hidden actions. One rendition allows agents to produce assets of low quality, or that are outright worthless, as when through history coins were clipped or notes counterfeited (Sargent and Velde 2003; Mihm 2007; Fung and Shao 2011). Li et al. (2012) have a fixed supply of one-period-lived assets yielding $\rho$. At some fixed cost $\kappa > 0$, agents can produce counterfeits that yield $0$. Assume counterfeits are confiscated by the government after each round of DM trade, so they do not circulate across periods, similar to Nosal and Wallace (2007). Then with $\theta = 1$, as above, the offer satisfies $c(q) \leq \rho d$ and the IC is

$$-(\phi - \phi^*)a + \beta \alpha \sigma \theta [u(q) - d\rho] \geq -\kappa + \beta \alpha \sigma \theta u(q),$$

where $\phi^* = \beta \rho$. The LHS of (48) is the payoff to accumulating genuine assets,

\textsuperscript{60}If an asset is risky but buyers and sellers have the same information when they meet, it functions well as a medium of exchange, with $\rho$ replaced by $E\rho$ in our benchmark model. But if agents see the realization of $\rho$ before DM trade, risk is reflected in the previous CM price, lowering the liquidity premium. Andolfatto and Martin (2013) and Andolfatto et al. (2014) assume the asset is risky and information about $\rho$ is available prior to DM trading at no cost. Nondisclosure (keeping information secret) is generally desirable, because trade based on expected $\rho$ better smooths consumption, obviously related to recent discussions of opacity and informational insensitivity.
the holding cost plus the DM surplus, while the RHS is the payoff to accumulating counterfeits, cost \( \kappa \) plus DM utility.

Given \( a = d \), in equilibrium, (48) can be rewritten as an upper bound on the amount of asset that can be transferred, as in Kiyotaki-Moore models,

\[
d \leq \frac{\kappa}{\phi - \phi^* + \beta \alpha \sigma \theta \rho}. \tag{49}
\]

This endogenizes bound increases with the cost of counterfeiting \( \kappa \), while it decreases with the cost of holding assets \( \phi - \phi^* \) and the frequency of DM trading opportunities \( \alpha \sigma \). If \( \alpha \sigma = 1 \), so there are no search or matching frictions in the DM, (49) says the asset’s value \( \phi d \) must be less than the cost of fraud \( \kappa \). As \( \kappa \to 0 \) the asset stops circulating. Notice an increase in the (endogenous) price \( \phi \) tightens the constraint, with policy implications described in Li et al. (2012). If (49) binds and \( d < A \) then the asset is illiquid at the margin and \( \phi = \phi^* \); if (49) does not bind the asset is perfectly liquid. There is an intermediate case where the liquidity premium, \( \kappa/A - \beta \alpha \sigma \theta \rho \), increases with \( \kappa \) and decreases with \( \beta \alpha \sigma \theta \). Notice the threat of fraud can affect asset prices even if there is no fraud in equilibrium, and assets with identical yields can be priced differently.\(^{61}\)

Given sellers may be uninformed about the quality of buyers’ assets, Lester et al. (2012) let them pay a cost to become informed. This endogenizes the fraction of matches where an asset is accepted (Kim 1996 and Berentsen and Rocheteau 2004 similarly let agents pay a cost to better recognize the quality of goods). To simplify the analysis, Lester et al. assume fraudulent assets are worthless and can be produced at no cost: \( \rho_L = \kappa = 0 \). This implies sellers only accept assets if they recognize them, and so it is possible to use standard bargaining theory – when sellers recognize assets there is no private information; when they don’t the assets are simply not on the bargaining table.

Interpret \( \sigma \) as the probability of a single coincidence times the probability a seller

\(^{61}\)If the asset subject to fraud is fiat money, as in Li and Rocheteau (2009), then (49) becomes \( c(q) \leq \kappa/\beta (\lambda + \alpha \sigma \theta) \), another case where acceptability is not exogenous, but depends on the policy variable \( \lambda \). Other applications include Li and Rocheteau (2011), Li et al. (2012), Shao (2013), Williamson (2014) and Gomis-Porqueras et al. (2014). The last paper includes two currencies and study exchange rates. In general, there can be many assets each with cost of fraud \( \kappa_j \) and supply \( \rho_j \), with aggregate liquidity a weighted average \( A_j \) with weights depending asset characteristics, consistent with the notion of money suggested by, e.g., Friedman and Schwartz (1970).
can discern, and hence accept, a buyer’s assets. At the beginning of the DM, sellers choose \( \sigma \) at cost \( C(\sigma) \), satisfying the usual conditions. The decision of a seller to become informed, so that he can accept assets in the DM, is similar to a decision to enter market in the first place or the search intensity decision discussed above. In order to give them some incentive to invest, sellers must have some bargaining power, so we use Kalai’s solution with \( \theta \in (0,1) \). The FOC is

\[
C'(\sigma) = \alpha \theta [u(q) - c(q)],
\]

where as always \( v(q) = \min\{A\rho, v(q^*)\} \), with \( \rho \) the dividend on genuine assets. This equates the marginal cost of becoming informed to the expected benefit from being able to accept assets. As the marginal cost of information decreases, \( \sigma \) and \( \phi \) increase. With long-lived asset, Lester et al. (2012) show there are strategic complementarities between buyers’ asset demand and sellers’ information investment, and this can naturally generate multiple equilibria.

Suppose there are two short-lived assets, both yielding \( \rho \). Asset 1 is perfectly recognizable at no cost in a fraction \( \bar{\sigma} \) of all meetings while asset 2 requires an ex ante investment \( C(\sigma_2) \) to be acceptable in a fraction \( \sigma_2 \) of meetings. Thus, in an endogenous fraction \( \sigma_2 \) of matches buyers can use both assets, while in fraction \( \sigma_1 = \bar{\sigma} - \sigma_2 \) they can use only asset 1. The investment decision satisfies

\[
C'(\sigma_2) = \alpha \theta_0 \{[u(q_2) - c(q_2)] - [u(q_1) - c(q_1)]\}. \tag{50}
\]

The RHS of (50) is the seller’s benefit of being informed, the extra surplus from having a payment capacity of \( (A_1 + A_2) \rho \) instead of \( A_1 \rho \). If there is an increase in the supply of recognizable assets \( A_1 \), agents invest less in information, asset 2 becomes less liquid and \( \phi_2 \) falls. If the recognizable asset is fiat money, the acceptability of the other asset is affected by policy. At \( \iota = 0 \), we get \( q_1 = q_2 = q^* \) and agents stop investing – reflecting the old idea that the use of money saves information costs.

There is more to be done on information, but existing results help in understanding phenomena related to acceptability and pledgeability, as well as the coexistence of assets with different returns, as discussed in Section 9 in the context of money and bonds. Information theory is a natural and venerable notion to bring to bear on Hicks’ suggestion with which we begun this essay.\textsuperscript{62}

\textsuperscript{62} Other applications include Nosal and Rocheteau (2011), Zhang (2014), Lotz and Zhang
11 Generalized OTC Markets

Papers spurred by Duffie et al. (2005) maintain tractability by restricting $\alpha \in \{0, 1\}$; Lagos and Rocheteau (2009) relax this. In spirit if not detail, the idea is to do what Section 6 does for monetary theory: extend second-generation models to a more general yet still tractable framework. The new model captures aspects of illiquid markets like participants adjusting positions to reduce trading needs. For simplicity, suppose all trade goes through dealers, as in Section 5 with $\alpha_D > 0 = \alpha_I$.

Let $\alpha \in \mathbb{R}_+$, and assume $u_j(a)$ is the flow utility of an agent with preference type $j \in \{0, \ldots, I\}$. Each $I$ draws a preference type $j \in \{0, \ldots, I\}$ at Poisson rate $\omega_j$, with $\sum_i \omega_i = \omega$. When $I$ with preference $j$ contacts $D$, they bargain over the $a_j$ that $I$ takes out of the meeting, and a payment that includes $D$’s cost of the transaction $q(a_j - a)$, where $q$ is the real asset price in the interdealer market, plus a fee $\varphi_j(a)$.

The choice of $a_j$ solves $\max_{a_j} \{\bar{u}_j(a') - rqa'\}$, where

$$\bar{u}_j(a) \equiv \frac{(r + \eta) u_j(a) + \sum_k \omega_k u_k(a)}{r + \eta + \omega}$$

and $\eta = \alpha_D (1 - \theta_D)$ is the arrival rate adjusted for bargaining power. Also,

$$\varphi_j(a) = \frac{\theta_D [\bar{u}_j(a_j) - \bar{u}_j(a) - rq(a_j - a)]}{r + \eta}$$

is the intermediation fee. Equilibrium is given by desired asset positions $(a_i, \ldots a_I)$, the fee $\varphi_j(a)$, the $q$ that clears the interdealer market and the distribution $(n_{ij})$ satisfying the usual conditions. To focus on the implications for asset prices and measures of liquidity, consider $u_i(a) = \psi_i \log a$, where $\psi_1 < \psi_2 < \ldots < \psi_I$, and let $\bar{\psi} = \sum_j \omega_j \psi_j / \omega$. Then the post-trade position of type $j$ who just met a dealer is

$$a_j = \frac{(r + \eta) a_j^\infty + \omega \bar{a}}{r + \eta + \omega},$$

where $a_j^\infty = \psi_j / rq$ would be the investor’s demand in a frictionless market, and $\bar{a} = \bar{\psi} / rq$ is demand from $I$ with average valuation $\bar{\psi}$. In a frictional market, $I$ chooses holdings between $a_j^\infty$ and $\bar{a}$, with the weight assigned to $a_j^\infty$ increasing in $\eta$. 

Hence, frictions concentrate the asset distribution. As frictions decrease, $a_j \to a_j^\infty$ and the distribution becomes more disperse, but aggregate demand is unchanged. A message is that one should not expect to identify frictions by looking at prices alone. Trade volume is $\alpha_D \sum_j n_{ij} |a_j - a_i| / 2$, lower than a frictionless economy. It increases with $\eta$, capturing the idea that large volume characterizes liquid markets, where $I$ can switch in and out of positions easily. With unrestricted asset holdings, frictions affect volume through the extensive and intensive margins. If $\alpha_D$ increases, the number of investors who are able to trade rises, but the number who are mismatched with their current portfolio falls. Also, higher $\alpha_D$ shifts the distribution across desired and actual holdings in a way that increases volume. This effect is shut off if one $a \in \{0, 1\}$, which has different predictions for trade volume after changes in the microstructure of the market. In Duffie et al. (2005), with $a \in \{0, 1\}$, trade volume is independent of dealers’ market power; here volume decreases with $\theta_D$.\(^{63}\)

Another conventional measure of financial liquidity is the bid-ask spread or intermediation fee. In Duffie et al. (2005), an increase in $\alpha_D$ raises $I$’s value of search for alternative traders, so spreads narrow. With unrestricted asset holdings, spreads still depend on $\alpha_D$, but also on the extent of mismatch between asset positions and valuations. Hence, the relationship between the spread and $\eta$ can be nonmonotone: under reasonable conditions, one can show the spread vanishes as $\eta \to 0$ or as $\eta \to \infty$. In liquid markets $I$ has good outside options, and hence $\varphi$ is small; in illiquid markets, $I$ trades very little, so $a_j^\infty$ is close to $A$, and $\varphi$ is small. The model also predicts a distribution of transactions, with spreads increasing in the size of a trade, as well as varying with $I$’s valuation.\(^{64}\) The model can also be extended to allow heterogeneity across $I$ in terms of arrival rates or bargaining power, and those with higher $\eta$ trade larger quantities at a lower cost per unit.

Trading delays are an integral feature of the microstructure in OTC markets. The time it takes to execute a trade not only influences volume and spreads, but is often used directly as a measure of liquidity. Lagos and Rocheteau (2007, 2008b)

\(^{63}\)Branzoli (2013) estimates a variant of this model using data from the municipal bond market. He finds $\theta_D$ is sufficiently high to reduce trade volume by 65% to 70%.

\(^{64}\)The relationship between spread and trade size generally depends on details. Lester et al. (2013) find with competitive search and a Leontief matching technology, costs decrease with the size of the trade, in accordance with evidence from corporate bond markets.
endogenize $\alpha_D$ with entry by $D$, and derive some new predictions, including a change in the equilibrium set due to the nonmonotonicity mentioned above. The model can generate multiple equilibria: it may be illiquid because participation by $D$ is low given a belief that $I$ will only trade small quantities; and it is rational for $I$ to take conservative positions given long trading delays. Tight spreads are correlated with large volume and short delays across equilibria, and scarce liquidity can arise as a self-fulfilling prophecy. Subsidizing entry can eliminate this multiplicity. Even when equilibrium is unique, the model has novel predictions, like lower market power for dealers promoting entry and reducing delays by increasing trade size. Similarly, a regulatory reform or a technological innovation that gives $I$ more direct access to the market (e.g., Electronic Communication Networks) implies an increase in market liquidity and intermediated trade.

The model also provides insights on welfare in illiquid markets. At least when contact rates are exogenous, in Duffie et al. (2005) welfare is unaffected by $\theta_D$, which only affects transfers from $I$ to $D$. When $a$ is not restricted to $\{0, 1\}$, equilibrium is inefficient unless $\theta_D = 0$. Indeed, if $D$ captures any of the gains that $I$ gets from adjusting his portfolio, $I$ economizes on intermediation fees by choosing $a_j$ closer to $\bar{a}$, thus increasing mismatch. When $\alpha_D$ is endogenous, the equilibrium is generically inefficient, again related to Hosios (1990). Efficient entry requires $\alpha_D$ equal the contribution of dealers to the matching process, but efficiency along the intensive margin requires $\alpha_D = 0$. As in monetary models, those two requirements are incompatible, and the market is inefficient, although as in many other models, a competitive search version can deliver efficiency (Lester et al. 2013). Branzoli (2013) finds that the most effective way to promote trading activity is the introduction of weekly auctions where investors trade bilaterally.

When $\alpha_I = 0 < \alpha_D$, investors trade only with dealers who continuously manage positions in a frictionless market. As mentioned earlier, some markets are well approximated by this, while others are better represented by $\alpha_I > 0 = \alpha_D$, such as the Federal Funds market, where overnight loans are traded, typically without intermediation. Afonso and Lagos (2015a) model the Fed Funds market explicitly, providing another case where the traders in the model are interpreted as banks. They have
$a \in \mathbb{R}_+$ and trade bilaterally, so the state variable is a time varying distribution of asset holdings $F_t(a)$. Afonso and Lagos show existence and uniqueness, characterize the terms of trade, and address various positive and normative questions, including quantitative questions facing central banks. Afonso and Lagos (2015b) consider the special case $a \in \{0, 1, 2\}$, similar to some monetary models.

In other applications, Biais et al. (2014) give a reinterpretation where agents have continuous access to the market, but learn their valuations infrequently. They also provide a class of utility functions nesting Duffie et al. (2005) and Lagos and Rocheteau (2009). Lagos and Rocheteau (2008) allow investors to have both infrequent access to the market, where they are price-takers, and to dealers, where they bargain. Pagnotta and Philippon (2012) study marketplaces competing on speed, endogenizing the efficiency of the trading technology, and entry/investment decisions from an industrial organization perspective. Melin (2012) has two types of assets traded in different markets, one with search and one frictionless. Mattesini and Nosal (2013), Geromichalos and Herrenbreuck (2013), Lagos and Zhang (2013) and Han (2015) integrate generalized models of OTC markets with the monetary models in Section 6, which would seem an important step in the program.

One reason it is important is the following: Duffie et al. (2005) and many subsequent papers have something called a liquidity shock, which is presumably meant to capture a need to offload financial assets in favor of more liquid payment instruments, but is more rigorously interpreted as a change in the utility of consuming the fruit of a Lucas tree. The papers that integrate models of money and generalized OTC markets take this a level deeper by having a shock to the utility of a good that is acquired in the DM, where liquid assets are required for exchange, as in many of the specifications presented above. Then the reason for trying to sell off assets is not that you do not want fruit, but that you want something that is easier to get using a more liquid asset. This seems more realistic, and more elegant, and moreover it allows one to study many other issues in a rigorous way – e.g., the effects of monetary policy on OTC markets.

In particular, in Lagos and Zhang (2013) an asset $a$ called equity with dividend $\rho$ is held by $I$ with time-varying idiosyncratic valuations. There are gains from trade
in a from heterogeneous valuations, and I participates in an OTC market like the DM in our benchmark monetary model, but intermediated by D, who again has access to a frictionless interdealer market. In this market fiat money is essential as a medium of exchange, and as usual some mechanism like bargaining determines the terms of trade between I and D. The DM alternates with a frictionless CM where agents rebalance portfolios. Equilibrium entails a cutoff preference type such that I below this who meets D trades all his equity for cash and I above this trades all his cash for equity. When I has all the bargaining power, he trades equity for cash at the price in the interdealer market. If D has all the bargaining power, I trades at a price higher (lower) than the interdealer market if I wants to buy (sell) assets.

This implies a bid-ask spread determined by monetary policy, and details of the market structure, such as the speed at which I contacts D or bargaining power. A nonmonetary equilibrium always exists, and when \( \pi \) is large it is the only equilibrium. In nonmonetary equilibrium only I holds equity, since there is no trade in the OTC market, and the equity price is the expected discounted value of the dividend stream for the average I. Monetary equilibrium exists for lower \( \pi \). For \( \pi \) not much larger than \( \beta - 1 \), there is a unique stationary monetary equilibrium where D holds all the equity overnight, while I holds it intraday. The asset price is higher than in nonmonetary equilibrium, because OTC exchange props up the resale value of assets. As usual, the Friedman rule implements the efficient allocation, and real asset prices decrease with \( \pi \) because money is complementary with (used to purchase) a. With entry by D, Lagos and Zhang (2013) generate sunspot equilibria with recurrent episodes that resemble financial crises – when a sunspot shock hits, spreads spike, while volume, trading frequency, market-making activity and asset prices collapse. This is again driven by the self-referential nature of liquidity.

Another area where New Monetarist models can provide useful and practical advice, somewhat related to the above-mentioned work on Fed Funds, concerns payment and settlement systems in general. Central banks around the world oversee such systems, e.g., the Division of Reserve Bank Operations and Payment Systems oversees the Federal Reserve Banks as providers of financial services to depository institutions and fiscal agency services to the government. A cashless New Keynesian
framework, standard GE theory or reduced-form monetary models cannot provide much guidance to the regulators of these systems, where trade is better characterized as bilateral. See Freeman (1996) for an early effort at modeling this activity carefully. See Williamson and Wright (2010b, Sec. 5), Nosal and Rocheteau (2011, Ch. 9) and the papers cited in fn. 54 for models designed to address the issues, and McAndrews et al. (2011) for a survey of related work.

12 Conclusion

The literature summarized above covers much territory, including rudimentary theories of commodity money, variations on and contributions to the general search-and-bargaining literature, fully-articulated quantitative macro systems, and micro depictions of financial institutions trading in OTC markets, like the market for Fed Funds. The models differ in various ways, but share some basic features, including an attempt to take the microfoundations of exchange seriously. Is it worth the effort? Lucas (2000) says “Successful applied science is done at many levels, sometimes close to its foundations, sometimes far away from them or without them altogether.” This sounds reasonable, especially when there are important issues that we think we need to address for which the foundations are poorly developed. But that is not the case in monetary economics, as we hope to have demonstrated.

When embarking on research we all have to decide what distance from foundations makes us comfortable, and here we offer some opinions on this. We do not endorse models that impose on agents behavior that they do not like. An example is the imposition of a CIA constraint in the model of bilateral trade in Diamond (1984). If two agents meet and want to barter, who are we to preclude that? The situation is worse in CIA models that otherwise adhere to GE methods, where agents trade only along their budget lines, because when they are not trading with each other one cannot even ask how they might like to trade. At least in Diamond’s example, this can be addressed by specifications for specialization precluding barter, as in Kiyotaki and Wright’s (1991) version of his model, but one has to also preclude credit, which has something to do with commitment and memory issues, as Kocherlakota (1998) makes clear. A related example of impositions on agents is assuming that
they trade taking prices as given in situations where the prices are all wrong due to nominal or other rigidities. It is natural to take prices as given in Arrow-Debreu, where they efficiently allocate resources. If prices are all wrong, due to rigidities, however, wouldn’t the economy evolve to other ways of allocating resources? If not, in our opinion, this needs to be explained, not assumed.

One sometimes hears that devices like CIA constraints or nominal contracts are “no different” than limiting what agents can do by specifying the environment in particular ways. We disagree. Frictions like spatial or temporal separation, limited commitment and imperfect information are features best made explicit, so one can work out all of their implications. With such frictions, it is atypical to get the first best, but this does not justify impositions on behavior rather than on the environment. Is having agents unable to adjust portfolios (e.g., increase money balances) after they meet, or unable to find the right counterparty in the first place, on par with assuming sticky prices or CIA constraints? We think not, because in models with sticky prices or CIA constraints agents trade inefficiently after they meet, leaving gains from trade sitting right there on the table. It seems different to have inefficient decisions made before meetings, like underinvestment in liquidity or search, than to have agents ignore gains from trade conditional on a meeting. Still, one can think about ways to reallocate liquidity after meetings, say through banks, or to encourage efficient search and investments, say through competitive search or creatively designed mechanisms, as discussed above.

We do not endorse the use of models where fiat currency — or bonds or bank deposits — enter utility or production functions directly. From a theoretical perspective, deriving the value of assets endogenously may mean having to work harder to tackle some issues, but from a policy perspective, anything that deviates from this discipline is obviously subject to the Lucas critique. The appropriateness of assumptions depends on the issue at hand, of course, but it is hard to imagine why anyone would prefer reduced-form models, unless the alternatives were too hard. But there is nothing especially hard about the material presented above. To be clear, real assets may generally appear in production or utility functions — e.g., capital, housing and wine all belong there. But if they somehow serve to facilitate transactions, that
is worth modeling explicitly.65

Our position is not, and cannot be, based on models that strive for deeper microfoundations empirically outperforming relatively reduced-form models. Whatever primitives one adopts, including frictions and mechanisms, the same results (for a given set of observables) obtain if one starts with the value function having assets as arguments. In our benchmark model, \( V(a) = W(a) + \alpha \sigma[u(q) - v(q)] \), where \( q \) depends on \( a \), and one can take this \( V(a) \) as a primitive to get an observationally equivalent reduced-form model. It is not so easy, however, to come up with a good guess for \( V(a) \), or even for its properties, out of thin air. Recall that Aruoba and Chugh (2008), e.g., show homothetic utility over goods does not imply \( V(a) \) is homothetic, and this has implications for optimal policy. Also, key ingredients in \( V(a) \) are \( \alpha \) and \( \sigma \), capturing search and matching, and \( v(q) \), nesting various mechanisms. In some models the distribution \( F(a) \) is also an important element, as are pledgeability and acceptability, as functions of private information. How does one know how these features figure into the reduced-form without deriving it?

Moreover, these features provide new avenues of exploration for policy issues. On inflation, in particular, the models presented incorporate several effects: (i) inflation is a tax on real balances; (ii) bargaining can compound this wedge; (iii) so can search-and-matching frictions; (iv) distortions revolving around participation, search intensity and reservation trade decisions interact with the effect of inflation, and for some parameterizations they imply that some inflation can be desirable; and (v) distributional considerations can also imply that some inflation may be desirable. While (i) is clear from standard reduced-form models, (ii)-(v) are not.66

---

65 The distinction can be subtle. The OTC model of Duffie et al. (2005) and Lagos and Rocheteau (2009) has agents getting flow utility from assets. This can be taken at face value for some assets, like housing or fruit-bearing trees. Or, it can be interpreted as a reduced-form for various liquidity and hedging services in the case of financial securities, like municipal or corporate bonds (Duffie et al. 2007; Garleanu 2009). We view the latter interpretation as acceptable when the focus is not on why people trade assets, but on the consequences of trading in frictional markets. However, in our baseline alternating-market model this is less of a problem: assets can give off coupons or dividends in CM numeraire or, realistically, dollars. To further emphasize the point, assets in the baseline Kiyotaki-Wright model also directly generate utility. But rather than saying money enters utility functions, it seems more accurate to say that some goods in the utility function end up endogenously playing a role commonly ascribed to money.

66 We reiterate that while the constraints in some of our models “look like” CIA restrictions, they are in fact feasibility conditions. It is clear that in Kiyotaki-Wright (1989), e.g., traders...
Also, while a nondegenerate $F(a)$ may be important for some issues, because it can be a natural outcome of decentralized trade, and captures a fundamental tension between two roles of monetary policy – providing favorable returns on liquidity and addressing liquidity-sharing considerations (see Wallace 2014) – for some purposes it is appropriate to work with models with degenerate distributions (as in Shi 1997 or Lagos-Wright 2005). Again, this depends on the question at hand.

Other issues discussed include the wisdom of trying to reduce unemployment using inflation, and the interpretation of sticky prices. Search and bargaining are not critical for making the first point qualitatively, and bargaining is replaced by posting for the second. Still, Berentsen et al. (2011) argue that both search and bargaining are quantitatively important in accounting for unemployment and inflation, and search is obviously the key to sticky prices in Head et al. (2012). Both search and bargaining are relevant for the quantitative effects of inflation. Lagos (2010b) shows how frictions help us understand issues in financial economics quantitatively, while Rocheteau and Wright (2013) argue that these kinds of models are consistent with outcomes that appear anomalous from the perspective of standard asset-pricing theory. The approach also sheds new light on inside vs outside money (Cavalcanti and Wallace 1999a,b), banking (Berentsen et al. 2007; Gu et al. 2013a), investment (Shi 1999a,b; Aruoba et al. 2011), OTC financial markets (Duffie et al. 2005; Lagos and Rocheteau 2009), and both conventional and unconventional monetary policy (Williamson 2012; Rocheteau et al. 2014), to mention a few additional examples.

At the frontier, research is trying to further develop models with multiple assets having different returns, without ad hoc restrictions on their use in transactions. This “modified Hahn problem” (Hellwig 1993) is challenging, but there are currently explanations under discussion: (i) certain pairwise trading mechanisms can deliver this as an outcome (Zhu and Wallace 2007); (ii) so can private information about asset quality (Li et al. 2012; Lester et al. 2012); (iii) so can assumptions about safety, e.g., from theft (He et al. 2008; Sanches and Williamson 2010); (iv) it can also cannot turn over something they do not have, but this cannot be called a CIA constraint because the baseline model does not even have cash. Moreover, the transactions pattern is endogenous and may not be unique. We reject the idea that the approach is “the same as” CIA or MUF models, despite the obvious result that one can always reverse engineer a reduced form to “look like” a microfounded model (e.g., as in Camera and Chien 2013).
be a self-fulfilling prophecy (Lagos 2013a); and (v) it is sometimes socially efficient (Kocherlakota 2003; Hu and Rocheteau 2013). Are these explanations satisfactory? Which are most relevant? While progress has been made, these are still important open questions. Another direction is to pursue qualitative and quantitative models combining New Monetarist and Keynesian features, as in examples by Aruoba and Schorfheide (2011) and Williamson (2015). More quantitative work on all this would be welcome. So would further research on banks and other intermediaries as providers of liquidity. Gu et al. (2015) suggest more work is needed on the theories that combine money and credit.\textsuperscript{67}

We close by highlighting a few issues where the methods covered above are especially useful or provide novel insights. First, they deliver endogenous exchange patterns that illustrate the interplay between intrinsic characteristics (e.g., storability or recognizability) and beliefs in determining which objects will or should be used to facilitate transactions. They determine the effective supply of liquidity, depending on the environment and policy. They allow us to study monetary, credit and intermediary arrangements, and allow us to clarify the essential frictions. Importantly, the approach is amenable to implementation theory and mechanism design, mapping frictions like commitment or information problems into incentive-feasible allocations, and identifying institutions with good welfare properties. This is problematic in reduced-form models, where the frictions are not well specified. Also, the framework is flexible enough to accommodate a variety of market structures. This is relevant for understanding how, e.g., the impact of inflation depends on whether the terms of trade are determined by bargaining, price posting or price taking, whether search is random or directed, etc. Future work should continue to explore different micro market structures.

The framework also illustrates how economies where liquidity considerations matter can be prone to multiplicity and interesting dynamics, where endogenous

\textsuperscript{67}They prove that in many natural environments, improvements in credit conditions are irrelevant in monetary equilibrium, because simply crowd out real balances one for one. This is very much like other irrelevance results (e.g., Modigliani-Miller or Ricardian equivalence), in that there may well be exceptions to the baseline results, as Gu et al. (2015) discuss, but in many standard models the results hold, and more generally, the results of changes in credit conditions can be very different in monetary and nonmonetary economies.
transaction patterns are not unique nor stationary. This is true of simple first- and second-generation models, as well as the more sophisticated versions designed to study the macro economy and financial markets. We mentioned how elements like $\alpha, \sigma, v, F$ and $\chi$ open up new avenues of exploration, e.g., in discussion of the hot potato effect, where velocity depends on explicit search, entry and trading decisions, or in the discussion of pledgeability and acceptability based on information theory. The models are also set up to analyze credit in the context of bilateral relationships, or pairwise meetings, where private information is naturally accommodated. Sometimes the models generate novel perspectives on topical issues, as with the zero lower bound problem in models with Nash bargaining, where $\nu < 0$ would be desirable but is not feasible.

The models can also contribute to discussions informational sensitivity, liquidity traps and price bubbles. The theory can be used to better understand the impact of OMO’s and less conventional policies. It can be used to deliver a fully-exploitable Phillips curve and sticky nominal prices, both of which are commonly discussed in macro, but usually with different implications. The theory can be used not only to make conceptual points about these issues, but also to organize and interpret micro data, e.g., as in the studies of sticky prices discussed above. It can also deliver time series observations where after a monetary injection, quantities first rise, then later prices rise, without sticky-price assumptions, but with prices and quantities determined bilaterally efficiently. The models generate a demand for liquidity that helps understand correlations between money holdings and nominal rates or bond holdings and spreads, without sticking assets in utility. Finally, we mention how the models allow one to analyze many dimensions of liquidity in a unified framework, including acceptability, pledgeability, moneyness, velocity, trade volume, bid-ask spreads and liquidity premia. We hope our presentation of this approach will be useful, especially given current interest in liquidity in economics and finance.
Appendix on Notation

\( \alpha, r, \beta = \) arrival rate, discount rate, discount factor

\( \rho = (\text{utility of}) \) dividend if \( \rho > 0 \) or storage cost if \( \rho < 0 \)

\( a, A = \) individual, aggregate asset holdings

\( \sigma, \delta = \) single- and double-coincidence prob

\( \kappa, \varepsilon = \) cost and probability of entry a la Pissarides

\( n_j = \) measure of type \( j \), \( m_i = \) measure inventory \( i \)

\( \mu = \) monitoring probability

\( n = \) fraction of monitored agents in CW

\( \tau = \) trading strategy

\( \tau = Y (\bar{t}) = \) best response correspondence

\( u, c = \) utility, cost of DM good

\( q, Q = \) quantity in monetary, barter trades

\( V^A, V^B, V^C, V^D = \) value fn for autarky, barter, credit and deviation

\( V_a = \) value fn for \( a \in \{0, 1\} \)

\( \theta_a = \) bargaining power of agent with \( a \in \{0, 1\} \)

\( v(q) = \) cost of \( q \) – i.e., a general mechanism

\( \eta = \alpha \theta = \) arrive rate times bargaining power

\( P, M, C = \) producer, middleman and consumer in RW

\( I, D = \) investor and dealer in DGP

\( S_I = \) surplus in DGP

\( \omega_j = \) probability of preference shock \( j \) in DGP

\( F(a) = \) asset dist’n

\( V(a), W(a) = \) value fn for \( a \in \mathbb{R}_+ \) in DM, CM

\( \phi, z = (\phi + \rho) a = \) price and value of \( a \)

\( U(x) - \ell = \) CM utility

\( S(q) = u(q) - c(q) = \) DM surplus

\( q, q_b = \) DM quantity in money and in barter trades

\( d, p = d/q = \) DM dollars and price

\( \pi, \iota = \) money growth (or inflation) and nominal interest rate

\( G, T = \) gov’t consumption and transfers

\( \lambda = \) liquidity premium or Lagrange multiplier

\( w, w_k = \) factor prices for labor and capital

\( t, t_k = \) tax rates on labor and capital income

\( \Delta = \) depreciation rate on \( k \)

\( \gamma, \Gamma = \) DM utility fn, \( \Gamma q^{1-\gamma} / (1 - \gamma) \)

\( D, \chi = \) KM debt and haircut parameter

\( \xi_b, \xi_b = \) wedges on shares and bonds in Lagos Journal of Monetary Economics

\( R = 1 + \iota = \) gross returns

\( \psi = \) DM preference shock, \( \psi u(q) \)

\( \zeta, 1 - \zeta = \) probability that \( \rho = \rho_H, \rho = \rho_L \)
Appendix on Commodity Money

We first derive $V_{12}$ from Section 2. Given dividends are realized next period, it should be simple to understand

$$(1 + r)V_{12} = \rho_2 + \alpha n_1 V_{12} + \alpha n_2 m_2 \left[ \tau_1 V_{13} + (1 - \tau_1) V_{12} \right] + \alpha n_2 (1 - m_2) (u + V_{12})$$

$$+ \alpha n_3 m_3 \left[ \tau_3 (u + V_{12}) + (1 - \tau_3) V_{12} \right] + \alpha n_3 (1 - m_3) V_{12}.$$

The RHS is type 1’s payoff next period from the dividend, plus the expected value of: meeting his own type with probability $n_1$, which implies no trade; meeting type 2 with their production good with probability $n_2 m_2$, which implies trade with probability $\tau_1$; meeting type 2 with good 1, which implies trade for sure; meeting type 3 with good 1, which implies trade with probability $\tau_3$; and meeting type 3 with good 2, which implies no trade. Algebra leads to (2).

Now to explain the SS condition, consider type 1 and pure strategies. If 1 has good 2, he can switch to good 3 only by trading with a 2 that has good 3 (since 3 never has good 3). For this, 1 has to meet 2 with good 3, which occurs with probability $n_2 m_2$, then trade, which occurs with probability $\tau_1$ (since 2 always wants good 2). And if 1 has good 3, he switches to good 2 only by acquiring good 1, consuming and producing a new good 2 (he never switches from good 3 to 2 directly, since if he preferred good 2 he would not trade it for good 3 in the first place). He trades good 3 for good 1 either by trading with 3 that has good 1, which occurs with probability $n_3 m_3$, or with 2 that has good 1 but prefers good 3, which occurs with probability $n_2 (1 - m_2) (1 - \tau_2)$, but we can ignore that since either $\tau_2 = 1$ or $m_2 = 1$. Equating the measure of type 1 that switch from good 2 to good 3 and the measure that switch back, we get (1).

<table>
<thead>
<tr>
<th>case</th>
<th>$\tau$</th>
<th>$m$</th>
<th>existence?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1, 1, 1)</td>
<td>(1/2, 1/2, 1/2)</td>
<td>never</td>
</tr>
<tr>
<td>2</td>
<td>(1, 1, 0)</td>
<td>($\sqrt{2}/2$, $\sqrt{2} - 1$, 1)</td>
<td>maybe</td>
</tr>
<tr>
<td>3</td>
<td>(1, 0, 1)</td>
<td>($\sqrt{2} - 1$, 1, $\sqrt{2}/2$)</td>
<td>never</td>
</tr>
<tr>
<td>4</td>
<td>(1, 0, 0)</td>
<td>(1/2, 1, 1)</td>
<td>never</td>
</tr>
<tr>
<td>5</td>
<td>(0, 1, 1)</td>
<td>(1, $\sqrt{2}/2$, $\sqrt{2} - 1$)</td>
<td>maybe</td>
</tr>
<tr>
<td>6</td>
<td>(0, 1, 0)</td>
<td>(1, 1/2, 1)</td>
<td>maybe</td>
</tr>
<tr>
<td>7</td>
<td>(0, 0, 1)</td>
<td>(1, 1, 1/2)</td>
<td>never</td>
</tr>
<tr>
<td>8</td>
<td>(0, 0, 0)</td>
<td>(1, 1, 1)</td>
<td>never</td>
</tr>
</tbody>
</table>

Table 1: Candidate Equilibria in the Commodity Money Model
Table 1 lists candidate equilibria, with existence results for the case $n_j = 1/3$. Consider case 1. After inserting $m$, the BR conditions reduce to

\[
\begin{align*}
\Delta_1 &\geq 0, \text{ or } \rho_3 - \rho_2 \geq u/6 \\
\Delta_2 &\geq 0, \text{ or } \rho_1 - \rho_3 \geq u/6 \\
\Delta_3 &\geq 0, \text{ or } \rho_2 - \rho_1 \geq u/6.
\end{align*}
\]

Since these cannot all hold, this is never an equilibrium. In contrast, for case 2, the BR conditions reduce to

\[
\begin{align*}
\Delta_1 &\geq 0, \text{ or } \rho_3 - \rho_2 \geq (1 - \sqrt{2})u/3 \\
\Delta_2 &\geq 0, \text{ or } \rho_1 - \rho_3 \geq 0 \\
\Delta_3 &\leq 0, \text{ or } \rho_2 - \rho_1 \leq (1 - \sqrt{2}/2)u/3.
\end{align*}
\]

For some parameters, these all hold and this is an equilibrium. The rest are similar.

\[\blacksquare\]
References


S. Boragan Aruoba and Frank Schorfheide (2011) “Sticky Prices vs Monetary Fric-
tions: An Estimation of Policy Tradeoffs,” *AJMacro* 3, 60-90.
Aleksander Berentsen, Guido Menzio and Randall Wright (2011) “Inflation and


Gabriele Camera and Yiting Li (2008) “Another Example of a Credit System that Coexists with Money,” *Journal of Money, Credit, and Banking* 40, 1295-308.


Edward J. Green and Warren E. Weber (1996) “Will the New $100 Bill Decrease Counterfeiting?” *FRB Minneapolis QR.*


Angela Kui Huang (2015) “Smaller or Larger Banks?” mimeo


95


John Law (1705) *Money And Trade Considered.*
Benjamin Lester, Andrew Postlewaite and Randall Wright (2011) “Information and Liquidity,” *Journal of Money, Credit, and Banking* 43, 355-76.
Victor Li (1994) “Inventory Accumulation in a Search-Based Monetary Economy,” *Journal of Monetary Economics* 34, 511-36.
nomic Dynamics 11, 388-404.


Ed Nosal and Guillaume Rocheteau (2013) “Pairwise Trades, Asset Prices, and
Monetary policy,” *Journal of Economic Dynamics and Control* 37, 1-17.


Guillaume Rocheteau, Randall Wright and Sylvia X. Xiao (2014) “Open Market
Operations,” mimeo.


Steven Williamson (2013) “Scarce Collateral, the Term Premium, and Quantitative Easing,” mimeo.
Steven Williamson and Randall Wright (2010a) “New Monetarist Economics: Methods,” FRB St. Louis Review.


