Money, intermediation, and banking ☆

David Andolfatto [a,b], Ed Nosal [c,*]

[a] Department of Economics, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6
[b] Rimini Centre for Economic Analysis, Italy
[c] Financial Markets Group, Federal Reserve Bank of Chicago, 230 South LaSalle Street, Chicago, IL 60604-1413, USA

ARTICLE INFO

ABSTRACT

The business of money creation is conceptually distinct from that of intermediation. Yet, these two activities are frequently—but not always—combined together in the form of a banking system. We develop a simple model to examine the question: When is banking essential? There is a role for money due to a lack of record-keeping and a role for intermediation due to the existence of private information: both money and intermediation are essential. When monitoring costs associated with intermediation are sufficiently low, the two activities can be separated from one another. However, when monitoring costs are sufficiently high, a banking system that combines these two activities is essential.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

We view a banking system as an institutional structure that combines two primary activities: liquidity provision and intermediation. We define liquidity provision as the supply of payments instruments and intermediation as the pooling, monitoring, and transformation of individual securities into alternative debt instruments. These two activities are conceptually distinct; a point highlighted by Friedman (1960), among others. In both theory and practice, liquidity provision can occur in the absence of intermediation and intermediation can occur in the absence of liquidity provision. Nevertheless, we frequently observe these two activities combined within agencies that are commonly called "banks" or institutional structures known as "banking systems." That is, banks are unique among intermediated structures in that their liabilities are designed to be money. To the best of our knowledge, we are unaware of any theory that explains why banking—the way in which we define it here—might be *essential*.[1]

There is, of course, a large literature on the theory of banking. But for most of this literature, the term *bank* can be replaced with *intermediary*. For example, Diamond and Dybvig (1983) explain the emergence of intermediated structures. But as the liabilities of their intermediaries do not circulate, they do not fit our definition of a bank. In Cavalcanti and Wallace (1999) banks issue money that circulates in the economy, but they are not intermediaries. Berentsen et al. (2007) examine a monetary economy in which a bank accepts deposits of cash and issues liabilities against cash reserves. However, in their model too, bank liabilities do not circulate. Similarly, in Smith (2003), banks are indistinguishable from, say, insurance companies as their liabilities do not circulate in any meaningful sense. Cavalcanti et al. (1999, 2005) examine models in which a subset of private agents issue liabilities backed by reserves of cash held at a clearinghouse. These agents resemble banks in that they accept cash deposits and issue liabilities that circulate. But their banks are not intermediaries; and hence, do not fit our definition of a bank. The same observation applies to the banks studied in He et al. (2005).

Naturally, our review of the literature above should not be construed as criticism, as most of these papers do not focus on the question of whether a banking system is essential. To address the question that concerns us here, we take the following approach. First, we ignore any meaningful role for fiat money. While fiat money plays an important role in modern economies—and arguably, in highly primitive ones—it seems clear that the development of banking and private money preceded the widespread use of government-issued fiat money (Hicks, 1989). Instead, we adopt a simple finite-horizon model, similar to Kiyotaki and Moore (2002), to introduce the need for a circulating medium. Moreover, to distinguish the business of banking from insurance, we abstract from insurance motives entirely by assuming risk-neutral agents. Intermediation in our model is instead motivated by the need for a delegated monitor, along the lines of Diamond (1984) and Williamson (1986).

Our model emphasizes two key frictions, both of which feature prominently in the literatures on money and intermediation. The first friction is the absence of a public-access record-keeping technology. The second friction is private information over the returns that are realized across individual investment projects. In the absence of either friction, the economy functions perfectly well without anything that one might label as money or as a banking system. Absent a record-keeping device, a tangible medium of exchange is essential. When private information is introduced and monitoring is costly, an intermediated structure—a delegated monitor—is essential. It is not, however, entirely obvious why the services of liquidity provision and intermediation should be performed by the same agent (or agency). Our main result is that even when money and intermediation are both essential, banking is inessential when monitoring costs are sufficiently small. However, a banking system is essential when monitoring costs are sufficiently large.

2. The benchmark environment

There are three dates, three time-dated goods and three types of agents, all labeled $i = 1, 2, 3$. There are N agents of each type, where $N > 1$ is a finite integer. A type-i agent prefers good i to good $i - 1$ (modulo 3), but receives no benefit from consuming good $i + 1$ (modulo 3).[2] Preferences are linear and given by

$$c_i + \varepsilon c_{i-1},$$

where c_i represents good i consumption and $0 < \varepsilon < 1$.

Output is divisible and non-storable. A type-i agent produces good $i - 1$ at date $i - 1$. Production outcomes are random at the individual level, where agent i's output realization—success or failure—is revealed at date $i - 1$. There is no aggregate uncertainty. In particular, at each date F agents fail to produce output, where $0 < F < N$. From an individual perspective, F/N represents the probability of failing to produce output. In the event of success, an agent produces $y > 0$ units of output. Since there is no aggregate uncertainty, total output at each date is given by $(1 - \lambda)Ny$, where $\lambda \equiv F/N$.

Throughout, we assume that type-1 agents have a commitment technology and that all other agents types do not. In the benchmark environment, we assume that individual outputs are observable and that there is a public-access record-keeping device. Later on, we modify this environment by first removing the public-assess record-keeping device, and then by assuming that individual outputs are private information. When output is assumed to be private information—in Section 5—we will introduce a monitoring technology.

3. The ex ante efficient allocation

Given our simple setup, the *ex ante* efficient allocation should be obvious. Since type-i agents value date i goods more than their own, a planner would allocate all date i output $(1 - \lambda)Ny$ to type-i agents. Because agents are risk-neutral, each type-i agent is indifferent between mechanisms that, in expectation, deliver $(1 - \lambda)y$ units of good i to him. Hence, an *ex ante* efficient allocation allows each agent to achieve an *ex ante* utility payoff equal to $(1 - \lambda)y$.

The implementation scheme described below builds on the fact that: (1) type-1 agents can commit; (2) output is observable; and (3) that there is a public-access record-keeping device. There are potentially many ways to implement the efficient allocation. Here is one. The allocation of good i is given by the following simple rule: the total amount of output

---

[2] In what follows, we will suppress the "modulo 3" qualification.

surrendered to a mechanism[3] at date 1 is equally divided among all type-1 agents; the total output surrendered to the mechanism at dates 2 and 3 is divided among the type-2 and type-3 agents, respectively, in proportion to output surrendered at dates 1 and 2. Agents play the following strategies: type-2 and type-3 agents surrender their output, if they have it, at dates 1 and 2, respectively; and, type-1 agents *promise* to surrender their output, if they have it, at date 3.

It is easy to see that the above allocation rule and strategies constitute an equilibrium. If all other agents play the proposed strategies, a type-2 or type-3 agent who defects by consuming his own output receives a payoff of εy, which is less than the equilibrium payoff of y. And, since type-1 agents can commit, they will surrender their output at date 3.

4. Lack of record-keeping

The allocation rule and strategies described above are infeasible if a public-access record-keeping device does not exist, since they rely on some form of public memory. For trade to occur, some sort of physical and non-counterfeitable object is needed (Kocherlakota, 1998).

Standard monetary models feature an infinite horizon, a complete lack of record-keeping and no commitment. In these environments, trade is facilitated by objects that are fiat in nature, such as intrinsically useless and unbacked tokens. In contrast, our environment has a finite horizon with some limited commitment in the form of type-1 agents' ability to commit. Here, monetary exchange can work off the fact that a subset of agents can commit to redeem tokens. As we shall see below, whether these tokens are created by society and endowed to type-1 agents, or whether type-1 agents create tokens on their own is irrelevant in terms of implementable allocations. All that is necessary is that some monetary object exist.

Imagine, then, that each type-1 agent is endowed with a divisible, durable, and non-counterfeitable token. Consider the following allocation rules and strategies. The allocation rule for output is similar to that described above: the total amount of output surrendered to the mechanism at date i is divided among the type-i agents in proportion to the amount of tokens they surrender. The allocation rule for tokens is: at date i the total amount of tokens surrendered to the mechanism is divided among the type-$i + 1$ agents in proportion to the output they surrendered. Agents play the following strategies. The strategy for tokens is: type-i agents surrender tokens at date i if they have them. The strategies for output are identical to the record-keeping environment: type-2 and type-3 agents surrender their output, if they have it, at dates 1 and 2, respectively; and, type-1 agents *promise* to surrender their output, if they have it, at date 3.

The allocation rules and strategies have been constructed in a way that makes it transparent that, in equilibrium, the supply of tokens serves as a perfect substitute for the missing public-access record-keeping device. In particular, if any type-2 or type-3 agent consumes his output, then he will be unable to consume any of the date 2 or date 3 goods, respectively.

Before we proceed, some brief remarks on the fiat-like nature of money are in order. While the tokens in the equilibrium described above do not have any *explicit* backing, they are not really pure fiat instruments since type-1 agents commit to accept them as payment for their output. In this sense, tokens are *de facto* backed by some commitment power. Therefore, instead of endowing agents with tokens, we could alternatively assume that type-1 agents create them, with an explicit promise of to redeem them for output in the future if they have it, and then use the above strategies and allocation rules to implement an *ex ante* efficient allocation.[4] Under this interpretation, tokens look a lot like inside money.[5]

5. Private information

We now assume that each agent's production outcome is private information and that type-1 agents can commit only to what is publicly observable.[6] In this situation, the strategies and allocation rules described in the previous section no longer constitute an equilibrium. In particular, type-1 agents will always have an incentive to that claim their output is

---

[3] A mechanism simply accepts and distributes output (and possibility other objects) according to a prescribed rule.

[4] From a date-3 perspective, tokens issued by type-1 agents will be heterogenous; some tokens are worth y—those issued by successful type-1 agents—and while others are worthless—those issued by unsuccessful type-1 agents. But given the allocation rules for tokens and goods, described above, the mechanism treats all type-1 money symmetrically. (We assume that agents surrendered their tokens before the output realizations are observed.) Implicitly, one can interpret the symmetric treatment of date-1 tokens as a form of intermediation. For example, suppose that at the beginning of date 3, all type-3 agents holding tokens, which are claims issued by type-1 agents, collectively pool them and issue new claims or tokens against them that give the holder proportional share of total output; this is done prior to the revelation of the type-1 agents' date 3 output realizations. Type-3 agents surrender output to the mechanism, type-3 agents (collectively) surrender the tokens issued by type-1 agents, and the mechanism distributes output in proportion to the type-3 tokens surrendered by type-3 agents. The act of the type-3 agents pooling the tokens issued by type-1 agents and then re-issuing new "riskless" tokens is the form of intermediation envisioned by Diamond (1984) and Williamson (1986). Note, however, that there is nothing fundamental about having a set of the type-3 agents acting as the intermediary; in principle, type-1 or type-2 agents may also perform this task.

[5] The notion of inside money here is similar to that in Cavalcanti and Wallace (1999). In Cavalcanti and Wallace (1999) agents who are monitored can issue inside money. In our setup, we can re-interpret type-1 agents as being monitored and assume that there exists a court that enforces contracts on the based on observables. Furthermore, since type-2 and type-3 agents are not monitored, they are not subject to the enforcement mechanism.

[6] If type-1 agents were able to commit unconditionally, then the introduction of private information has no effect on equilibrium outcomes. The assumption that (type-1) agents can only commit to what is observable is standard in the literature, e.g., Townsend (1979) and Williamson (1986).

zero—since output is not publicly observable—and those who actually produce y will consume it. Without any further modifications to the environment, the only equilibrium is autarky.

Following Townsend (1979) and Williamson (1986), we introduce a costly monitoring technology. The technology works as follows. If a producing agent fails to surrender his output, then the actual level of production—0 or y—can be revealed if he is monitored. The cost of monitoring a producing agent is $\mu \geqslant 0$ utils. The monitoring cost can be shared or spread out over any number of agents. If, for example, M agents each expend μ/M utils, then their combined effort allows them to monitor one producing agent. Individual monitoring costs, as well as the monitoring outcome, are observable.

It turns out that monitoring is relevant only at date 3, and that only type-1 and type-3 agents would ever have an incentive to monitor.[7] Type-3 agents may have an incentive to monitor because, in any equilibrium with trade, they are holding tokens (or inside money) that can be used to purchase the output they value. Note that because type-3 agents have no commitment power, their decision to monitor must be sequentially rational. In contrast, since type-1 agents can commit, their decision to monitor (themselves) need not respect sequential rationality.

We will define an intermediary as a set of agents who collectively perform the task of monitoring.[8] Since type-2 agents do not have an incentive to monitor, they cannot be an intermediary.

5.1. Type-3 intermediation

In this section, we consider strategies that can be interpreted as giving rise to an institutional structure that separates money-issue and intermediation. More specifically, we assume that type-1 agents are responsible for creating the economy's monetary instrument and type-3 agents are collectively responsible for intermediation. In terms of the timing of events for type-3 intermediation, we assume that the decision to monitor or not is made *after* date-3 output is surrendered (by type-1 agents) and distributed (to type-3 agents). This timing serves to emphasize the fact that type-3 agents cannot commit to future actions.

Consider the following equilibrium strategies and allocation rules, most of which are identical to those described in the previous section. The allocation rule for goods at date i is to divide the total amount of goods surrendered to the mechanism at date i among the type-i agents in proportion to the amount of tokens they surrendered. The allocation rule for tokens divides the total amount of tokens surrendered to the mechanism at date i among the type-$i + 1$ agents in proportion to the amount of output surrendered at date i. The strategies for goods for type-i agents to surrender output if they have it. In addition, (i) at date 1, each type-1 agent creates a token that is redeemable for y units of date 3 output, if he has it; and (ii) successful type-3 producers who surrender tokens at date 3 monitor all type-1 agents who did not surrender output at date 3, if total date-3 output surrendered is less than $(N - F)y$; otherwise they do not monitor anyone.

The interesting questions to examine here are: (1) will successful type-1 producers surrender their output at date 3? and (2) will type-3 agents monitor if total date 3 output falls below $(N - F)y$? Since we are interested in implementing a truth-telling equilibrium, let us suppose that all successful type-1 agents surrender their output at date 3, and then examine whether, conditional on this behavior, any individual type-1 agent has an incentive to withhold his output. Note that such a deviation will be profitable if there is even the smallest chance that the deviating type-1 agent escapes monitoring. Therefore, any equilibrium requires that if total output is less than $(N - F)y$, then any agent who does not surrender output is monitored with probability one.[9]

Suppose then that a successful type-1 agent does not surrender his output at date 3. The above strategies imply that type-3 agents with tokens will monitor all of the $F + 1$ type-1 agents who did not surrender output. In this case, the cost of monitoring is $\mu(F + 1)/(N - F)$ utils for each type-3 agent who monitors. This monitoring activity will recover y units of the "hidden" output.

Whether type-1 agents have an incentive to surrender output depends on whether the delegated monitors finds it sequentially rational to monitor in the event of a defection. Let U_3^m denote the expected payoff to monitoring and U_3^n denote the expected payoff to not monitoring. Then,

$$U_3^m = y - \mu \left(\frac{F + 1}{N - F} \right)$$

and

$$U_3^n = \left(\frac{N - F - 1}{N - F} \right) y.$$

---

[7] A type-2 agent would never monitor at date 3 since he does not value date-3 output and cannot commit to monitoring.

[8] In footnote 4, we motivated the symmetric treatment of tokens, issued by type-1 agents, at date 3 as a form of intermediation. Continuing this line of discussion, in an environment characterized by asymmetric information, if a group of agents transforms risky tokens issued by type-1 agents into riskless ones at date 3, they must also be willing to perform, or at least manage, the task of monitoring in the event that total output surrendered falls short of $(1 - \lambda)Ny$. So, the set of agents that are involved in transforming claims are also the set of agents who must manage monitoring. Since, without loss of generality, we assume that the mechanism treats all date-1 issued claims symmetrically, the set of agents who do the monitoring *is* the intermediary.

[9] Implicitly, we assume that sequential monitoring of type-1 agents is not possible. Allowing for this changes a bit of the arithmetic, but not our main conclusions.

Hence, type-3 agents will decide to monitor if $U_3^m \geqslant U_3^n$, and $U_3^m \geqslant U_3^n$ if and only if

$$y \geqslant \mu(F + 1) = \mu(\lambda N + 1).$$

This inequality is obviously satisfied if $\mu = 0$; but in general, we can define a reservation monitoring cost μ^* such that $y \equiv \mu^*(\lambda N + 1)$ or

$$\mu^* = \left(\frac{y}{\lambda N + 1}\right). \tag{1}$$

We can conclude that monitoring by type-3 agents with money will be sequentially rational for any $\mu \in [0, \mu^*]$.[10] Hence, if μ is sufficiently small in this sense, then conditional on all other successful type-1 producing agents surrendering their output, a successful type-1 producing agent will not have an incentive to hide his output. In other words, the *ex ante* efficient allocation can be implemented under an institutional arrangement that separates the businesses of the money-issue from intermediation. Hence, a banking arrangement is not essential. If, however, $\mu > \mu^*$, then the *ex ante* efficient allocation cannot be implemented under such an arrangement.

5.2. Type-1 intermediation: banking

Since type-1 agents can commit, the monitoring/intermediation function can simply be delegated collectively to them. That is, when the type-1 agents create their tokens (or liabilities) at date 1, they also commit to monitoring all type-1 agents who do not surrender output at date 3, in the event that aggregate output falls short of $(N - F)y$. Since this threat of monitoring is credible, no type-1 agent has an incentive to hide output at date 3; any hidden output will ultimately be discovered and confiscated. Hence, in equilibrium, no monitoring will occur and the *ex ante* efficient allocation can be implementable regardless of the size of μ.

In the stark environment that we consider, a banking arrangement weakly dominates an institutional arrangement that separates the businesses of money-issue and intermediation over the entire parameter space and strictly dominates for parameter configurations (y, μ, λ, N) that satisfy $y < \mu(\lambda N + 1)$. It is in these latter cases that we say that banking is essential.

It is interesting to note that μ^* is a decreasing function of N, see (1). It is tempting to interpret N as a measure of population size. If this is the case, then the model has a nice implication that banking arrangements are likely to be more prevalent in larger economies (*ceteris paribus*, of course). The model also suggests that banking arrangements are likely to be more prevalent in environments where monitoring costs are high and/or the probability of failure is high, relative to the return to investment. The parameter μ reflects the difficulty of acquiring information—say, in a bankruptcy proceeding—and there is some reason to believe that this parameter may increase over time as economic relationships and accounting practices grow in complexity.

6. Discussion

Our model is able to capture some, but certainly not all, aspects of the development of money and banking in history. First, our model is consistent with the idea that anyone with a capacity to issue money will endeavor to do so to meet the demand for liquidity. Absent legal restrictions prohibiting the practice of small note issue, this type of behavior is prevalent throughout recent history. For example, Bodenhorn (1993) quotes an Italian General Secretary of the Banco D'Italia how, prior to 1874, "everyone was issuing notes, even individuals and commercial firms; the country was overrun with little notes of 50, 25, and 20 centimes issued by everyone who liked to do so." The author also notes that when state legislation banned U.S. banks from issuing notes of less than \$5, railroad companies, public houses, merchants and even churches filled the void with their own notes. Even Smith (1937, pp. 305–313) noted, with some disapproval one might add, how small notes drove specie from the country.

One aspect of history that our model cannot account for is the widespread regulatory efforts that were expended to prohibit private small note issue. One common argument was that holders of a few small notes had little incentive to expend real resources on monitoring or determining authenticity. Even those holding large quantities of notes may have had little incentive to monitor, since others might have incentive to free-ride on their efforts. But as Bodenhorn (1993, p. 822) reports, in antebellum America, at least, independent businessmen known as note brokers made markets in banknotes; at the same time providing information and monitoring services. This institutional setup bears some resemblance to the one that emerges in our model where money-issue is separated from monitoring.

What, then, of the emergence of banking arrangements? The interpretation that we offer here is not with respect to the emergence of agencies often referred to as banks *per se*; but rather, the emergence of what one might call a banking system, broadly defined to include regulatory agencies. As noted by Klein (1974), many early U.S. banks became members of private certifying and monitoring agencies, which performed some of the functions similar to modern central banks. A famous example is the Suffolk banking system, e.g., see Smith and Weber (1999). In an interesting paper, Gorton and Mullineaux

---

[10] Suppose only a subset of type-3 agents with tokens decide to monitor after they observe the amount the amount of output is $(N - F - 1)y$. Owing to the linearity of preferences, the condition for monitoring for any subset of the type-3 agents with tokens is also given by condition (1).

(1987) argue that the capacity of private note brokers to monitor and control the behavior of bank managers was increasingly eroded as demand deposits came to supplant bank notes during the nineteenth century. While our model is not rich enough to distinguish between bank notes and demand deposits, this historical development can be thought of manifesting itself as an increase in our model parameter μ, and hence, an increase in the likelihood of the emergence of banking.

7. Conclusion

We describe an environment where, absent record-keeping, money is essential. Whether this money takes the form of a fiat token or of an explicitly backed private liability is indeterminate, although we do argue that a fiat token in conjunction with type 1 agents' promises to surrender output amounts to inside money. Either way, however, a specific subset of agents must initially be in possession of the monetary instrument; whether it is endowed to them, or whether they create it themselves, is irrelevant.

When the environment is further modified by assuming that idiosyncratic production shocks are private information and that such information can only be revealed through costly monitoring, some form of intermediation is essential. This is to say that it becomes essential for some group of agents to agree collectively monitor the issuers of some pool of securities. But it is not obvious if it matters who these agents are. If monitoring costs are sufficiently small—in the sense that $\mu \leqslant \mu^*$—then the monitoring function can be delegated either to those who issue the money or to those who are ultimately in a position to redeem it. On the other hand, if monitoring costs are sufficiently high—in the sense that $\mu > \mu^*$—then it is essential that the monitoring function is delegated to those agents responsible for creating the economy's monetary instrument.

While our model is very simple, it arguably allows us to interpret some aspects of the development of money and banking in recent history. As such, the basic ideas embedded within the model may serve to develop a richer class of models designed to explain more complicated aspects of the way payments systems are organized and regulated.

References

Berentsen, A., Camera, G., Waller, C., 2007. Money, credit and banking. Journal of Economic Theory 127, 171–195.

Bodenhorn, H., 1993. Small-denomination banknotes in antebellum America. Journal of Money, Credit, and Banking 25, 812–827.

Cavalcanti, R., Wallace, N., 1999. Inside and outside money as alternative media of exchange. Journal of Money, Credit, and Banking 31, 443–457.

Cavalcanti, R., Erosa, A., Temzelides, T., 1999. Private money and reserve management in a random matching model. Journal of Political Economy 107, 929–945.

Cavalcanti, R., Erosa, A., Temzelides, T., 2005. Liquidity, money creation and destruction, and the returns to banking. International Economic Review 46, 675–706.

Diamond, D., 1984. Financial intermediation and delegated monitoring. Review of Economic Studies 51, 393–414.

Diamond, D., Dybvig, P., 1983. Bank runs, deposit insurance and liquidity. Journal of Political Economy 91, 401–419.

Friedman, M., 1960. A Program for Monetary Stability. Fordham University Press, New York.

Gorton, G., Mullineaux, D., 1987. The joint production of confidence: endogenous regulation and nineteenth century commercial-bank clearinghouses. Journal of Money, Credit, and Banking 19, 457–468.

He, P., Huang, L., Wright, R., 2005. Money and banking in search equilibrium. International Economic Review 46, 637–670.

Hicks, J., 1989. A Market Theory of Money. Clarendon Press, Oxford.

Klein, B., 1974. The competitive supply of money. Journal of Money, Credit, and Banking 6, 423–453.

Kiyotaki, N., Moore, J., 2002. Evil is the root of all money. American Economic Review 92, 62–66.

Kocherlakota, N., 1998. Money is memory. Journal of Economic Theory 81, 232–251.

Smith, A., 1937. An Inquiry into the Nature and Causes of the Wealth of Nations. The Modern Library, New York.

Smith, B., 2003. Taking intermediation seriously. Journal of Money, Credit, and Banking 35, 1319–1357.

Smith, B., Weber, W., 1999. Private money creation and the suffolk banking system. Journal of Money, Credit, and Banking 31, 624–659.

Townsend, R., 1979. Optimal contracts and competitive markets with costly state verification. Journal of Economic Theory 21, 265–293.

Williamson, S., 1986. Costly monitoring, financial intermediation and equilibrium credit rationing. Journal of Monetary Economics, 159–179.