Liquidity, asset prices and systemic risk

Franklin Allen, Department of Finance, Wharton School,
University of Pennsylvania
Douglas Gale, Department of Economics, New York University[1]

Abstract

Central banks often intervene in the financial system to prevent crises. They frequently cite financial fragility or the contagion that might otherwise occur as a justification for their actions. This argument has traditionally been based on historical experience rather than a theoretical understanding of these phenomena. This paper discusses a theoretical framework for considering these issues and the role of central bank intervention.

1. Introduction

In August 1998 the Russian government defaulted on its domestic debt. Despite the fact that the amount of this debt was small relative to the total value of assets in the world, the event had a large effect on the global financial system. On the day the default was announced, three quarters of the stock markets in the world fell (Kaminsky and Schmukler (1999)). In the weeks that followed, there was considerable turbulence in financial markets. In October, the Federal Reserve Bank of New York facilitated a takeover of the hedge fund Long-Term Capital Management (LTCM) by a consortium of major banks. The unusual movements in asset prices following the Russian default had brought LTCM to the brink of bankruptcy. If LTCM had gone into bankruptcy, its assets would have been liquidated. The precise way in which this bankruptcy would have been handled was fraught with uncertainty. LTCM was incorporated in the Cayman Islands and there were few precedents for this type of event there (Allen and Herring (2001)).

One rationale for the New York Fed's action was the argument that the financial markets in which LTCM traded were fragile and subject to contagion: rapid liquidation of such a substantial amount of assets would have overwhelmed the liquidity available in the markets, causing a significant drop in asset prices. This would have caused problems for other intermediaries that in turn might have been forced to liquidate assets, causing prices to fall even further. The cumulative effect of LTCM's default might have been a global financial crisis. The New York Fed's action pre-empted this possibility and markets soon stabilised.

Did the bankruptcy of LTCM really pose a systemic risk for the global financial system? Would asset prices have collapsed if LTCM had been forced to liquidate its assets in a short space of time? Many have doubted this and argued that the New York Fed acted inappropriately. Standard models of asset pricing suggest that a single liquidation of the size of LTCM will not lead to a meltdown in asset prices. According to these models, asset prices are determined by the discounted stream of cash flows generated by the assets. Changes in the supply of assets to the market does not affect their price provided that such changes do not signal information (Scholes (1972)). It seems unlikely that LTCM's bankruptcy would have signalled very much about the future cash flows of corporations or discount rates in the global economy. So, according to this view, the New York Fed's intervention was unnecessary.

In this paper we review recent theories of financial crises. In particular, we are interested in understanding the systemic risk associated with financial fragility and contagion and how central banks should respond. The rest of the paper is organised as follows. Section 2 reviews the early literature on

[1] Allenf@wharton.upenn.edu, Douglas.Gale@nyu.edu.

financial crises, based on the Diamond-Dybvig model of bank runs. Section 3 introduces a more recent class of models in which the source of financial crises is real business cycle shocks. This section focuses on the welfare economics of financial crises from the point of view of optimal risk-sharing. One of the important elements of this discussion is the relationship between market provision of liquidity and its effect on asset prices, which is further explored in Section 4. In Section 5 we return to the debate about whether financial crises result from real business cycle shocks or self-fulfilling expectations. Section 6 discusses models of contagion. Section 7 sums up our discussion of the policy implications of the research reviewed here.

2. Risk-sharing

First-generation models of financial crises were developed in the 1980s, beginning with seminal work on bank runs by Bryant (1980) and Diamond and Dybvig (1983) (hereafter DD). Important contributions were also made by Chari and Jagannathan (1988), Chari (1989), Champ et al (1996), Jacklin (1987), Jacklin and Bhattacharya (1988), Postlewaite and Vives (1987), Wallace (1988; 1990) and others. DD is at the core of most models in the literature on bank-centred financial crises. A typical DD-style model has three dates $t = 0, 1, 2$ and a large number of identical consumers. Each consumer is endowed with one unit of a homogeneous consumption good. At date 1, the consumers learn whether they are early consumers, who only value consumption at date 1, or late consumers, who only value consumption at date 2. Consumers' uncertainty creates a preference for liquidity.

An individual can invest in a combination of long-term (illiquid) investments, yielding high returns, and short-term (liquid) assets, yielding low returns. The short-term asset pays a return of one unit after one period and the long-term asset pays a return $r < 1$ after one period or $R > 1$ after two periods. The long asset has a higher return if held to maturity, but liquidating it in the middle period is costly, so it is not very useful for providing liquidity. An individual investor faces a difficult choice between return and liquidity. If he holds the long asset and turns out to be an early consumer, he will lack liquidity. If he holds the short asset and turns out to be a late consumer, his returns will be low. What he really wants is insurance against his uncertain demand for liquidity, but this he cannot provide by holding a mixture of the two assets.

Banks have a comparative advantage in providing liquidity insurance. The bank can offer each depositor a superior contract, promising a combination of liquidity and high returns that an individual investor cannot match using markets. For simplicity, assume that the fraction of early consumers is constant. Thus, there is no uncertainty about the aggregate demand for liquidity. There is only uncertainty about which individuals will demand liquidity at the intermediate date. At the first date, consumers deposit their endowments in the banks, which invest them on behalf of the depositors. In exchange, depositors are promised a fixed amount of consumption at each subsequent date, depending on when they choose to withdraw. The banking sector is perfectly competitive, so banks offer risk-sharing contracts that maximise depositors' ex ante expected utility, subject to a zero profit constraint.

DD provides a simple explanation of bank runs. The optimal insurance scheme requires the bank to promise depositors a fixed payment if they withdraw early. If too many depositors withdraw, the bank is unable to meet its promises without liquidating assets. Under some conditions, if most or all of the early depositors withdraw early, there will be nothing left for those who withdraw late. Thus, a bank run becomes a self-fulfilling prophecy: if a depositor believes that others will withdraw their deposits from the bank, it becomes optimal for the depositor to withdraw his deposits too.

There are two Nash equilibria of this "game", one in which only the early consumers (those who have received a liquidity shock) withdraw early and one in which everyone withdraws early. The former is incentive-efficient, the latter is not. What determines which equilibrium is observed? Market psychology? Animal spirits? Sunspots? We return to this point in Section 5 below.

Here we want to emphasise the importance of DD as a contribution to the microeconomic theory of intermediation. Apart from its usefulness as a model of bank runs, the DD model is remarkable for providing an account of what banks do and why they are needed. The insurance function (converting liquid liabilities into illiquid assets) is interesting in its own right, because it provides a theory of banking based on rational optimising behaviour and opens it up to microeconomic analysis. The same approach can be extended to the welfare analysis of monetary and banking policy.

3. Optimal financial crises

Gorton (1988) finds evidence from the United States during the National Banking Era which is consistent with the view that banking panics are related to the business cycle. The five worst recessions, as measured by the change in pig iron production, were accompanied by panics. In all, panics occurred in seven out of the 11 cycles. Using the liabilities of failed businesses as a leading economic indicator, Gorton finds that panics were systematic events: whenever this leading economic indicator reached a certain threshold, a panic ensued. The stylised facts uncovered by Gorton thus suggest banking panics are intimately related to the state of the business cycle. Calomiris and Gorton (1991) consider a broad range of evidence and reach similar conclusions.

A number of authors have developed models of banking panics caused by aggregate risk. Wallace (1988; 1990), Chari (1989) and Champ et al (1996) extend Diamond and Dybvig (1983) by assuming that the fraction of the population requiring liquidity is random. Chari and Jagannathan (1988), Jacklin and Bhattacharya (1988), Hellwig (1994) and Alonso (1996) introduce aggregate uncertainty which can be interpreted as business cycle risk. Chari and Jagannathan (1988) focus on a signal extraction problem where part of the population observes a signal about future returns. Others must then try to deduce from observed withdrawals whether an unfavourable signal was received by this group or whether liquidity needs happen to be high. Chari and Jagannathan are able to show panics occur not only when the outlook is poor but also when liquidity needs turn out to be high. Jacklin and Bhattacharya (1988) also consider a model where some depositors receive an interim signal about risk. They show that the optimality of bank deposits compared to equities depends on the characteristics of the risky investment. Hellwig (1994) considers a model where the reinvestment rate is random and shows that the risk should be borne both by early and late withdrawers. Alonso (1996) demonstrates using numerical examples that contracts where runs occur may be better than contracts which ensure runs do not occur because they improve risk-sharing.

Starting from this point of view, Allen and Gale (1998) (hereafter AG) construct a model in which financial crises are driven by fundamentals. An economic downturn will reduce the value of bank assets, raising the possibility of banks being unable to meet their commitments. If depositors receive information about an impending downturn, they will anticipate financial difficulties in the banking sector and try to withdraw their funds. This attempt will precipitate a crisis.

The main objective of AG is to analyse the welfare properties of the model and understand the role of central banks in dealing with panics. Bank runs are an inevitable consequence of the standard deposit contract in a world with aggregate uncertainty about asset returns. Furthermore, they play a useful role insofar as they allow the banking system to share these risks among depositors.

The basic assumptions about technology and preferences have become standard in the literature since the appearance of DD. AG retains many of the standard assumptions in the model but it differs from DD in important ways.

- There are aggregate shocks to asset returns. More precisely, banks hold long-term (illiquid) assets that earn a random return $\tilde{R}$ at date 2. Moreover, the returns to the risky assets are perfectly correlated across banks. Uncertainty about asset returns is intended to capture the impact of the business cycle on the value of bank assets. Information about returns becomes available before the returns are realised and when the information is bad it has the power to precipitate a crisis.
- The long-term asset is completely illiquid: none of the returns from this asset are available for sharing out among the early consumers. This is different from assuming that $r = 0$ in the DD model. Here the long asset cannot be touched in the short run. The "winding-up" of an insolvent bank must take time and, more importantly, there will be something left for late withdrawing depositors.
- AG does not impose the first come, first served assumption. (This assumption has been the subject of some debate in the literature as it is not an optimal arrangement in the basic DD model (see Wallace (1988) and Calomiris and Kahn (1991)). Instead, an insolvent bank shares its liquid assets equally among the early withdrawers. Those who do not withdraw early have to wait to obtain their funds and, again, they share the remaining assets equally.

In a number of countries and historical time periods, banks have had the right to delay payment for some time period on certain types of account. This is rather different from the first come, first served assumption. Sprague (1910) recounts how in the United States in the late 19th century people could obtain liquidity once a panic had started by using certified cheques. These cheques traded at a discount.

In the simplest version of the model, which serves as a benchmark for the rest of AG, there are no costs of early withdrawal, apart from the potential distortions that bank runs create for equilibrium risk-sharing and portfolio choice. In this context, a laissez faire banking system which is vulnerable to crises can actually achieve the first-best allocation of risk and investment. The first-best allocation can be identified with an optimal mechanism-design problem in which the allocation can be made contingent on a leading economic indicator (eg the return on the risky asset), but not on the depositors' types. By contrast, a standard deposit contract cannot be made contingent on the leading indicator. However, depositors can observe the leading indicator and make their withdrawal decisions contingent on it. When late-consuming depositors observe that returns are high, they can deduce that they will obtain more by waiting and are content to leave their funds in the bank until the last date. When the returns are low, they deduce that they are better off to withdraw their funds than leave them in, causing a bank run. The somewhat surprising result is that the optimal deposit contract produces the same portfolio and consumption allocation as the first-best allocation. The possibility of equilibrium bank runs allows banks to hold the first-best portfolio and produces just the right degree of contingency to provide first-best risk-sharing.

The idea that financial crises can be optimal is an important one. It is often taken as axiomatic among policymakers and in the literature that crises should be avoided at all costs. As the example in AG indicates, crises can perform a useful role in sharing risk. In fact, Allen and Gale (2000a) are able to show in the context of a more general model of banking and financial markets that crises can be constrained-efficient in a wide range of circumstances. They argue that policies towards crises need to be based on a careful understanding of the nature of the market failure that occurs. In the absence of market failure, intervention by the central bank may not be justified.

To provide a rationale for central bank intervention, some cost of early liquidation has to be introduced. AG considers a second version of the model in which the storage technology available to the banks is strictly more productive than the storage technology available to the late consumers who withdraw their deposits in a bank run. In these circumstances, where crises are costly, appropriate central bank intervention can avoid the unnecessary costs of bank runs while continuing to allow runs to fulfil their risk-sharing function. A bank run, by forcing the early liquidation of too much of the safe asset, actually reduces the amount of consumption available to depositors. In this case, laissez faire does not achieve the first-best allocation. This provides a rationale for central bank intervention. The central bank can intervene with a monetary injection that implements the first-best allocation. Suppose that a bank promises the depositor a fixed nominal amount and that, in the event of a run, the central bank makes an interest-free loan to the bank. The bank can meet its commitments by paying out cash, thus avoiding premature liquidation of the safe asset. Equilibrium adjustments of the price level at the two dates ensure that early and late consumers end up with the correct amount of consumption at each date and the bank ends up with the money it needs to repay its loan to the central bank. The first-best allocation is thus implemented by a combination of a standard deposit contract and bank runs.

Finally, AG considers the role of markets for the illiquid asset in providing liquidity for the banking system. The first two versions of the model have the very special feature that the risky, long-term asset is completely illiquid. There is no way of liquidating the risky asset to meet the claims of the early consumers, and this plays an important role in the "equilibration" of a bank run: the fact that some assets are always left over at the final date means that it can never be optimal in equilibrium for all the late consumers to join a run and withdraw early.

In the third version of the model, there is an asset market in which the risky asset can be traded and this provides a means of liquidating the long-term asset. Somewhat surprisingly, the introduction of asset markets leads to a Pareto *reduction* in welfare in the laissez faire case. The bankruptcy rules force the bank to liquidate as much of its assets as possible in an attempt to meet the claims of the depositors who withdraw early. Liquidation turns out to be self-defeating because the asset sales drive down the prices available on the market and the depositors are the losers. Once again, though, central bank intervention in the form of a monetary injection allows the financial system to share risks without incurring the costs of inefficient investment.

4. Market provision of liquidity

It is worth dwelling on the role of asset markets in some detail, since it has broad methodological implications for the analysis of crises. As the discussion of LTCM illustrates, the provision of liquidity to the market plays an important role in the analysis of financial fragility.

In both DD and AG, assets are represented by investment technologies. The short-term (liquid) asset is represented by a storage technology and the long-term (illiquid) asset is represented by a two-period investment technology. In the DD model, the possibility of premature liquidation of the long-term assets is also represented by a technology. If the long-term asset is liquidated prematurely, it yields a return of only $r < 1$ per unit invested. The difference in returns, $R - r$, is the cost of liquidation. The costly liquidation technology reflects the assumption that, when financial institutions have to realise the value of their assets in a hurry, they are typically unable to realise the full value that they would receive if they could wait until maturity. This loss of value is one of the costs of financial distress. However, the use of a reduced-form "liquidation technology" obscures a number of interesting features that are highly relevant for understanding the welfare economics of financial crises.

It is easy to see why the introduction of an asset market might amplify the effects of a bank run. By making all assets liquid, the new market allows the bank to liquidate all its assets in an attempt to meet the claims of the early withdrawers. Now the banks may be forced to liquidate their previously illiquid assets in order to meet their deposit liabilities. However, by selling assets during a run, they force down the price and make the crisis worse. This destroys the equilibrating mechanism of the earlier versions of the model in which the returns to the illiquid asset were untouchable at date 1.

Liquidation is obviously self-defeating, in the sense that it transfers value from depositors to the speculators in the market. A transfer is not inefficient and it might be thought that, unlike in DD, the premature liquidation does not involve a deadweight loss. The welfare analysis of the market's impact is a bit more subtle, however. The deadweight loss associated with liquidations takes the form of sub-optimal risk-sharing, not a loss of value per se. Optimal insurance would provide depositors with a positive transfer in bad states, where asset returns are low, and impose a tax or negative transfer in good states, where returns are high. The asset market does the opposite. By making transfers in the worst states, the market provides depositors with negative insurance.

In this case, there is an incentive for the central bank to intervene to prevent a collapse of asset prices, but again the problem is not runs per se but the unnecessary liquidations they promote. Another solution, explored in Allen and Gale (2000a), is the provision of optimal liquidity insurance through the market. Liquidity insurance takes the form of Arrow securities in theory and of credit derivatives in practice. If insurance markets are complete, banks can insure against runs and crises and once again achieve optimal risk-sharing. This is not to say that complete insurance eliminates crises - it may be socially optimal to have crises because of the flexibility default introduces into risk-sharing contracts - but simply that the market will determine the optimal incidence of financial crises. Conversely, incompleteness of insurance markets may provide a rationale for central bank intervention.

The role of liquidity in determining asset prices is explored in a different context by Allen and Gale (1994). However, the same feature that assets have to be sold at a loss in some states occurs there. When a liquidity shock hits the market, forcing some investors to sell assets quickly, there are two possible regimes. If the amount of liquidity in the market, as measured by holdings of liquid assets, is high, then the asset price is determined by the expected future returns to the asset. On the other hand, if the amount of liquidity in the market is low, then the price is determined by the amount of "cash in the market". Of course, the amount of liquidity is itself endogenous, and results from prior decisions by investors. Liquidity providers need a profit to induce them to participate in the market for assets. Speculators have an incentive to hold liquid assets in order to buy up assets only if the price is low enough. So, in some states, the market has to be illiquid and there has to be "cash-in-the-market" pricing.

In summary, modelling the provision of liquidity by the market instead of assuming banks have a costly "liquidation technology" is a methodological innovation in several respects:

- First, the cost of liquidation is now endogenous. Whether there is a loss of value in selling assets in the intermediate period is determined by the liquidity of the market, that is, by the portfolio choices of the investors and institutions that make up the market.
- Ex post, there is no deadweight loss from selling assets. An asset sale involves a transfer, but the asset's returns are not affected by the sale. This is a major change from the DD

model and its successors, in which the returns of the liquidated asset are determined by the technology and assumed to be lower than the asset's returns at maturity.

- Ex ante, liquidation may impose a cost. While the seller's "loss" is the buyer's "gain" ex post, they are both losers ex ante. Liquidation is inefficient ex ante because it does not provide the bank with insurance against changes in the asset price. The bank obtains a good price in states where the demand for liquidity is low and a bad price in other states where the demand for liquidity is high.
- The market's provision of liquidity is necessarily inadequate. Because the return on holding the short asset is lower than the return on holding the long asset, investors require some additional incentive for providing liquidity. They obtain this incentive in the form of a capital gain if they can buy the long asset cheaply in the middle period and realise a high return in the last period. This will happen only if there is a distress sale from which they can profit. In other words, the market will be willing to provide liquidity to a distressed institution only if the terms are sufficiently profitable and this means that the assets have to be sold "at a loss". Thus, the amount of liquidity provided in equilibrium will never be adequate to support asset prices at a level that would give optimal risk-sharing for banks.

5. Sunspots

Theoretical research on speculative currency attacks, banking panics, and contagion have taken a number of approaches. One is built on the foundations provided by early research on bank runs (eg Allen and Gale (1998; 2000a-d), Chang and Velasco (2000; 2001)) and Peck and Shell (1999)). Other approaches include macroeconomic models of currency crises that developed from the insights of Krugman (1979), Obstfeld (1986) and Calvo (1988) (see, for example, Corsetti et al (1999) for a recent contribution and Flood and Marion (1999) for a survey), game theoretic models (see Morris and Shin (1998; 2000) and Morris (2000) for an overview), amplification mechanisms (eg Cole and Kehoe (2000) and Chari and Kehoe (2000)) and the borrowing of foreign currency by firms (eg Aghion et al (2001)).

Two main perspectives on financial crises can be discerned in this literature. One is that they are *random events*, unrelated to changes in the real economy. The classical form of this view suggests that crises are the result of "mob psychology" or "mass hysteria" (see, for example, Kindleberger (1978)). The modern version, developed by DD and others, is that bank runs are self-fulfilling prophecies. As we saw in Section 2, there are two equilibria in the DD model, one with runs and one without. Which of these two equilibria occurs depends on extraneous variables or "sunspots". Although sunspots have no effect on the real data of the economy, they affect depositors' beliefs in a way that turns out to be self-fulfilling. (Postlewaite and Vives (1987) have shown how runs can be generated in a model with a unique equilibrium.)

The alternative to the sunspot view, discussed in Section 3, is that financial crises are a natural outgrowth of the *business cycle*. An economic downturn will reduce the value of bank assets, raising the possibility that banks are unable to meet their commitments. If depositors receive information about an impending downturn in the cycle, they will anticipate financial difficulties in the banking sector and try to withdraw their funds. This attempt will precipitate the crisis. According to this interpretation, panics are not random events but a response to unfolding economic circumstances. Mitchell (1941), for example, writes (p 74):

> "when prosperity merges into crisis ... heavy failures are likely to occur, and no one can tell what enterprises will be crippled by them. The one certainty is that the banks holding the paper of bankrupt firms will suffer delay and perhaps a serious loss on collection."

In other words, panics are an integral part of the business cycle.

Whichever view one takes of the causes of financial crises, there is some consensus based on historical experience that financial systems can be *fragile*. The threat of a financial crisis lies in the possibility that it will propagate through the economic system, causing damage disproportionate to the original shock. This notion of financial fragility is most easily seen in the sunspot model: the impact of extraneous uncertainty is equivalent to financial fragility, since the shock that "causes" the crisis has no effect on the fundamentals of the economy. Financial fragility can also be captured in a real business cycle model, where crises result from exogenous shocks. In this context, financial fragility is interpreted as a situation in which very small shocks can tip the economy over the edge into a full blown crisis. In other words, financial fragility is an extreme case of excess sensitivity to small shocks.

In terms of causation, the difference between sunspots (sensitivity to extraneous uncertainty) and excess sensitivity (extreme sensitivity to real exogenous shocks) is not great. The first could be thought of as a limiting case of the second. However, there are important differences between the two approaches. The sunspot theory does not predict crises; it allows for the possibility of crises. Furthermore, the sunspot theory also depends on fundamentals. Weak fundamentals are not sufficient for a crisis, but in the presence of weak fundamentals, self-fulfilling expectations may be sufficient for a crisis.

An approach that spans both the real business cycle approach and the sunspot theory is represented by AG, who call a crisis *essential* if, for certain parameter values, every equilibrium of the model is characterised by a crisis. Restricting attention to situations in which crises are essential gives the theory greater predictive power. These models may allow for sunspot equilibria, but do not rely on them.

A related approach is represented by the work of Morris and Shin (1998; 2000) and Morris (2000), who study models with multiple equilibria but use equilibrium selection arguments based on small amounts of asymmetric information about parameter values to predict which equilibrium will be chosen.

Using a special case of the framework developed in Allen and Gale (2000), Allen and Gale (2001) investigate the connection between financial fragility and the existence of sunspot equilibria. The connection is close. Financial fragility can occur when the spillover effect from liquidation of assets by banks is channelled to other banks through the price of assets in the market. What is crucial for understanding this phenomenon is the fact that the system minimises liquidity to be the minimum needed for preventing a crisis in certain states. If the demand for liquidity rises above this level, there will be a sharp fall in the price of assets. This drop in asset prices may force other banks into insolvency and exacerbate the crisis. The pecuniary externalities, to use the technical term, from one set of agents forces another much larger set into bankruptcy. In other words, a small shock (to liquidity demand) can have a large effect. In the limit, when the initial shock that causes the crisis becomes vanishingly small, we have something that looks very much like a sunspot equilibrium. However, the approach is different, since it does not rely on multiple equilibria.

The reason for financial fragility is the necessity for providing incentives to hold liquidity. It seems possible that as in AG the central bank can eliminate the inefficiencies associated with crises by an appropriate injection of money. This is an important topic for future research.

6. Contagion

The AG framework has also been used to construct a model in which small shocks lead to large effects by means of contagion - more precisely, in which a shock within a single sector can spread to other sectors and lead to an economy-wide financial crisis. Allen and Gale (2000b) construct a model in which, under certain circumstances, contagion is unavoidable when the economy is subject to a small shock.

The economy consists of a number of regions. The number of early and late consumers in each region fluctuates randomly, but the aggregate demand for liquidity is constant. This allows for inter-regional insurance as regions with liquidity surpluses provide liquidity for regions with liquidity shortages. One way to organise the provision of insurance is through an interbank market in deposits. Suppose that region *A* has a large number of early consumers when region *B* has a low number of early consumers, and vice versa. Since regions *A* and *B* are otherwise identical, their deposits are perfect substitutes. The banks exchange deposits at the first date before they observe the liquidity shocks. If region *A* has a higher than average number of early consumers at date 1, then banks in region *A* can meet their obligations by liquidating some of their deposits in the banks of region *B*. Region *B* is happy to oblige, because it has an excess supply of liquidity in the form of the short asset. At the final date the process is reversed, as banks in region *B* liquidate the deposits they hold in region *A* to meet the above average demand from late consumers in region *B*.

Inter-regional cross-holdings of deposits work well as long as there is enough liquidity in the banking system as a whole. If there is an excess demand for liquidity, however, the financial linkages caused by these cross-holdings can turn out to be a disaster. While cross-holdings of deposits are useful for reallocating liquidity within the banking system, they cannot increase the total amount of liquidity. If the economy-wide demand from consumers is greater than the stock of the short asset, the only way to provide more consumption is to liquidate the long asset. This is very costly (see Shleifer and Vishny (1992) and Allen and Gale (1998) for a discussion of the costs of premature liquidation), so banks try to avoid liquidating the long asset whenever possible. In this case, they can avoid liquidating the long asset by liquidating their claims on other regions instead. This mutual liquidation of claims does not create any additional liquidity, however. It merely denies liquidity to the troubled region and bank runs and bankruptcy may be the result. What begins as a financial crisis in one region can then spread by contagion to other regions because of the cross-holdings of deposits.

The interbank market works quite differently from the retail market. In the latter case, runs occur because deposit contracts commit banks to a fixed payment and banks must begin liquidating the long asset when they cannot meet liquidity demand from the short asset. In the interbank market the initial problem is caused by the fact that banks with an excess demand for liquidity cannot obtain anything from banks in other regions. This is the opposite of the problem in the retail market and, unlike there, cannot be solved by making the contracts discretionary or contingent since whatever their form they cancel each other out. Instead of being caused by the nature of interbank claims, spillovers and contagion result just from the fall in the value of bank assets in adjacent regions.

Whether the financial crisis does spread depends crucially on the pattern of interconnectedness generated by the cross-holdings of deposits. If the interbank market is *complete* and each region is connected to all the other regions, the initial impact of a financial crisis in one region may be attenuated. On the other hand, if the interbank market is *incomplete*, each region is connected with a small number of other regions. The initial impact of the financial crisis may be felt very strongly in those neighbouring regions, with the result that they too succumb to a crisis. As each region is affected by the crisis, it prompts premature liquidation of the long asset, with a consequent loss of value, so that previously unaffected regions find that they too are affected because their claims on the region in crisis have fallen in value.

It is important to note the role of the free rider problem in explaining the difference between a complete and incomplete interbank market. There is a natural pecking order among different sources for liquidity. A bank will meet withdrawals first from the short asset, then from holdings in other regions, and only in the last resort will it choose to liquidate the long asset. Cross-holdings are useful for redistributing liquidity, but they do not create liquidity; so when there is a global shortage of liquidity (withdrawals exceed short assets), the only solution is to liquidate long assets. If every region takes a small hit (liquidates a small amount of the long asset), there may be no need for a global crisis. This is what happens with complete markets: banks in the troubled region have direct claims on banks in every other region and there is no way to avoid paying one's share. With incomplete markets, banks in the troubled region have a direct claim only on the banks in adjacent regions. The banks in other regions pursue their own interests and refuse to liquidate the long asset until they find themselves on the front line of the contagion.

The notion of a region is intended as a metaphor for categories of banks that may differ in several dimensions. For example, some banks may be better at raising funds while other banks are better at lending them. Or it might be that banks focus on lending to different industries or in different regions and as a result have lending opportunities that are not perfectly co-related with their deposit base. In either case, an interbank market plays an important role in redistributing the funds efficiently. However, the existence of claims between different categories of banks opens up the possibility of contagion when one category is hit by a sudden demand for liquidity.

The reason that contagion can occur here is the existence of incomplete markets. The central bank can play an important role here by completing markets. If it is linked to all the banks, then it can overcome the free rider problem and simply reallocate liquidity to prevent the contagion.

7. Policy conclusions

The citing of "financial fragility" and "contagion" is often the rationale for intervention in the financial system by central banks and governments. Traditionally, the justification for intervention was based on historical evidence. The memory of the Great Depression and earlier crises is still with us and it powerfully reinforces the belief that such intervention is worthwhile. Until recently, there has been little attempt to try and understand these phenomena at a theoretical level. Although the state of the theory is too underdeveloped to allow for strong policy conclusions, some simple lessons can be drawn from the work reviewed here.

- In the first place, a micro-based theoretical analysis allows us to address *normative* questions about financial crises - for example, when are they consistent with optimal risk-sharing? - in addition to *positive* questions about what causes crises and how they can be prevented. Once the focus is on the welfare economics, we are led to think about the *optimality* of financial crises rather than mere crisis avoidance.
- A second lesson is that, in these models, the cost of financial crises comes from inefficient asset liquidation rather than the crisis per se. This may be because there is a real loss of asset value, as in DD, or because liquidation is associated with inefficient risk-sharing, as in AG. In either case, the policy recommendation is to avoid inefficient liquidation rather than prevent crises at all costs.
- There are several ways of avoiding the costs associated with inefficient liquidation. One is to substitute money for real claims, as in AG; another is to provide complete insurance through the market, as in Allen and Gale (2000a); another is to provide liquidity through the lender of last resort (LOLR), as in Bhattacharya and Gale (1987).
- Finally, we have seen that, under certain conditions, the laissez faire outcome is incentive-efficient or constrained-efficient, in which case there is no role for the LOLR. On the other hand, various frictions and imperfections give rise to the possibility that efficiency requires some intervention by the LOLR. For example, if insurance markets are incomplete, there may be a role for the central bank as a substitute for incomplete markets.

Our discussion has focused on financial issues, narrowly defined, and in particular on optimal risk-sharing. But it also has to be recognised that disruption of the financial sector has implications for the "real" sector (cf Bernanke and Gertler (1989)). The concern about financial fragility arises precisely because of the fear that what begins as a purely financial disturbance may spill over into the rest of the economy and cause a period of slow growth or even a recession. We have not discussed these issues explicitly, but we have examined models in which small shocks can have far-reaching consequences in the financial sector. We presume that when these disturbances do impose costs on the rest of the economy, there is a rationale for central bank intervention to prevent asset price volatility and bank defaults before they wreak havoc elsewhere.

Again, we have seen that these crises arise from mispricing of liquidity and/or lack of liquidity. For example, in AG, provision of liquidity by the market requires price volatility. The low return on liquid assets means that there must be states of the world where these can be used to make a profit. In Allen and Gale (2000b), a small shock in one region or sector can spread by contagion and cause a meltdown in the financial system if markets are incomplete. The discontinuity associated with bankruptcy means that even a small shock can have a large effect if it cascades sequentially through the financial system. In each case, liquidity provision by the LOLR appears to be the key to prevention.

Intervention by the central bank to provide liquidity is not the only way to deal with crises. Bank regulation such as capital controls is another instrument that can potentially be used to intervene. Bankruptcy law is another type of policy that is potentially important in controlling the effects of crises. In the models discussed, the bankruptcy law is such that banks must liquidate their assets to meet their obligations. Alternative laws that do not have this requirement but delay claims may be helpful in eliminating financial fragility and contagion. Much work remains to be done in the area of public policy and crises.

References

Aghion, P, P Bacchetta and A Banerjee (2001): "Currency crises and monetary policy in an economy with credit constraints", *European Economic Review* 45, 1121-50.

Allen, F and D Gale (1994): "Limited market participation and volatility of asset prices", *American Economic Review* 84, 933-55.

—— (1998): "Optimal financial crises", *Journal of Finance* 53, 1245-84.

—— (2000a): "Banking and markets", Working Paper 00-44B, Wharton Financial Institutions Center, University of Pennsylvania.

—— (2000b): "Financial contagion", *Journal of Political Economy* 108, 1-33.

—— (2000c): "Optimal currency crises", *Carnegie Rochester Series on Public Policy* 53, 177-230.

—— (2000d): "Bubbles and crises", *The Economic Journal* 110 (2000), 236-56.

—— (2001): "Financial fragility", Working Paper 01-37, Wharton Financial Institutions Center, University of Pennsylvania.

Allen, F and R Herring (2001): "Banking regulation versus securities market regulation", Working Paper 01-29, Wharton Financial Institutions Center, University of Pennsylvania.

Alonso, I (1996): "On avoiding bank runs", *Journal of Monetary Economics* 37, 73-87.

Bhattacharya, S and D Gale (1987): "Preference shocks, liquidity and central bank policy", in W Barnett and K Singleton (eds), *New approaches to monetary economics*, Cambridge University Press, New York.

Bernanke, B and M Gertler (1989): "Agency costs, net worth, and business fluctuations", *American Economic Review* 79, 14-31.

Bryant, J (1980): "A model of reserves, bank runs, and deposit insurance", *Journal of Banking and Finance* 4, 335-44.

Calomiris, C and G Gorton (1991): "The origins of banking panics, models, facts, and bank regulation", in Glenn Hubbard (ed), *Financial markets and financial crises*, University of Chicago Press, Chicago, IL.

Calomiris, C and C Kahn (1991): "The role of demandable debt in structuring optimal banking arrangements", *American Economic Review* 81, 497-513.

Calvo, G (1988): "Servicing the public debt: the role of expectations", *American Economic Review* 78, 1411-28.

Champ, B, B Smith and S Williamson (1996): "Currency elasticity and banking panics: theory and evidence", *Canadian Journal of Economics* 29, 828-64.

Chang, R and A Velasco (2000): "Financial fragility and the exchange rate regime", *Journal of Economic Theory* 92, 1-34.

—— (2001): "A model of financial crises in emerging markets", *Quarterly Journal of Economics* 116, 489-517.

Chari, V (1989): "Banking without deposit insurance or bank panics: lessons from a model of the US national banking system", *Federal Reserve Bank of Minneapolis Quarterly Review* 13 (Summer), 3-19.

Chari, V and R Jagannathan (1988): "Banking panics, information, and rational expectations equilibrium", *Journal of Finance* 43, 749-60.

Chari, V and P Kehoe (2000): "Financial crises as herds", working paper, Federal Reserve Bank of Minneapolis.

Cole, H and T Kehoe (2000): "Self-fulfilling debt crises", *Review of Economic Studies* 67, 91-116.

Corsetti, G, P Pesenti and N Roubini (1999): "Paper tigers? A model of the Asian crisis", *European Economic Review* 43, 1211-36.

Diamond, D and P Dybvig (1983): "Bank runs, deposit insurance, and liquidity", *Journal of Political Economy* 91, 401-19.

Flood, R and N Marion (1999): "Perspectives on recent currency crises literature", *International Journal of Finance and Economics* 4, 1-26.

Gorton, G (1988): "Banking panics and business cycles", *Oxford Economic Papers* 40, 751-81.

Hellwig, M (1994): "Liquidity provision, banking, and the allocation of interest rate risk", *European Economic Review* 38, 1363-89.

Jacklin, C (1987): "Demand deposits, trading restrictions, and risk-sharing", in E Prescott and N Wallace (eds), *Contractual arrangements for intertemporal trade*, University of Minnesota Press, Minneapolis, MN.

Jacklin, C and S Bhattacharya (1988): "Distinguishing panics and information-based bank runs: welfare and policy implications", *Journal of Political Economy* 96, 568-92.

Kaminsky, G and S Schmukler (1999): "What triggers market jitters? A chronicle of the Asian crisis", *Journal of International Money and Finance* 18, 537-60.

Kindleberger, C (1978): *Manias, panics, and crashes: a history of financial crises*, Basic Books, New York, NY.

Krugman, P (1979): "A model of balance-of-payments crises", Journal of Money Credit and Banking 11, 311-25.

Mitchell, W (1941): *Business cycles and their causes*, University of California Press, Berkeley, CA.

Morris, S (2000): "Contagion", *Review of Economic Studies* 67, 57-78.

Morris, S and H S Shin (1998): "Unique equilibrium in a model of self-fulfilling currency attacks", *American Economic Review* 88, 587-97.

—— (2000): "Global games: theory and applications", working paper, Yale University.

Obstfeld, M (1986): "Rational and self-fulfilling balance of payments crises", *American Economic Review* 76, 72-81.

Peck, J and K Shell (1999): "Bank portfolio restrictions and equilibrium bank runs", working paper, Cornell University.

Postlewaite, A and X Vives (1987): "Bank runs as an equilibrium phenomenon", *Journal of Political Economy* 95, 485-91.

Scholes, M (1972): "The market for securities: substitution versus price pressure and the effects of information on share price", *Journal of Business* 45, 179-211.

Shleifer, A and R Vishny (1992): "Liquidation values and debt capacity: a market equilibrium approach", *Journal of Finance* 47: 1343-66.

Sprague, O (1910): *A history of crises under the national banking system*, US Government Printing Office, Washington DC.

Wallace, N (1988): "Another attempt to explain an illiquid banking system: the Diamond and Dybvig model with sequential service taken seriously", *Federal Reserve Bank of Minneapolis Quarterly Review* 12 (Fall), 3-16.

—— (1990): "A banking model in which partial suspension is best", *Federal Reserve Bank of Minneapolis Quarterly Review* 14 (Fall), 11-23.