3

Intermediation and crises

With the exception of the Bretton Woods period from 1945 to 1971, banking crises have occurred fairly frequently in the past 150 years, as we saw in Chapter 1. There are two traditional ways of understanding crises. The first asserts that crises result from panics; the second asserts that crises arise from fundamental causes that are part of the business cycle. Both have a long history. For example, Friedman and Schwartz (1963) and Kindleberger (1978) argued that many banking crises resulted from unwarranted panics and that most of the banks that were forced to close in such episodes were illiquid rather than insolvent. The alternative view, put by Mitchell (1941) and others, is that financial crises occur when depositors have reason to believe that economic fundamentals in the near future look poor. In that case depositors, anticipating that future loan defaults will make it impossible for the bank to repay their deposits, withdraw their money now. The depositors in this case are anticipating insolvency rather than illiquidity. In this chapter, we consider both approaches.

Although the economic theory of banking goes back over 200 years, it was not until recently that a model of banking, in the contemporary sense, was provided in the seminal papers of Bryant (1980) and Diamond and Dybvig (1983). The publication of these papers marked an important advance in the theory of banking. Although the objective of the papers was to provide an explanation of bank runs, an equally important contribution was to provide a microeconomic account of banking activity that was distinct from other financial institutions. In fact, the papers contributed four separate elements to the theory of banking:

- a **maturity structure** of bank assets, in which less liquid assets earn higher returns;
- a theory of **liquidity preference**, modeled as uncertainty about the timing of consumption;
- the representation of a bank as an intermediary that provides **insurance** to depositors against liquidity (preference) shocks;
- an explanation of bank runs by depositors. In the case of Diamond and Dybvig (1983), the bank runs are modeled as the result of **self-fulfilling prophecies or panics**; in the case of Bryant (1980), they are modeled as the result of **fundamentals**.

In Sections 3.1–3.4 of this chapter, we describe a model of banking, loosely based on the Bryant (1980) and Diamond–Dybvig (1983) models, that shows how all these pieces fit together (see Chapter 2 for an introductory development of some of these ideas). Sections 3.5 and 3.6 develop a model based on the view that crises arise from panics while Section 3.7 develops a model based on the view that crises result from fundamentals. Section 3.8 contains a literature review and Section 3.9 concluding remarks.

3.1 THE LIQUIDITY PROBLEM

It is a truism that banks have liquid liabilities and illiquid assets. In other words, they borrow short and lend long. This makes banks vulnerable to sudden demands for liquidity (bank runs), but more on this later. This maturity mismatch reflects the underlying structure of the economy: individuals have a preference for liquidity but the most profitable investment opportunities take a long time to pay off. Banks are an efficient way of bridging the gap between the maturity structure embedded in the technology and liquidity preference.

We adopt the period structure introduced in Chapter 2. There are three dates indexed by $t = 0, 1, 2$. At each date there is a single good that can be used for consumption and investment and serves as a numeraire. There are two types of assets:

- The liquid asset (also called the *short asset*) is a constant returns to scale technology that takes one unit of the good at date t and converts it into one unit of the good at date $t + 1$, where $t = 0, 1$.
- The illiquid asset (also called the *long asset*) is a constant returns to scale technology that takes one unit of the good at date 0 and transforms it into $R > 1$ units of the good at date 2; if the long asset is liquidated prematurely at date 1 then it pays $0 < r < 1$ units of the good for each unit invested.

At the first date there is a large number, strictly speaking a continuum, of ex ante identical economic agents (consumers, depositors).[1] Each consumer has an endowment of one unit of the good at date 0 and nothing at the later dates. In order to provide for future consumption, agents will have to invest, directly or indirectly, in the long and short assets.

[1] We represent the set of agents by the unit interval $[0, 1]$, where each point in the interval is a different agent. We normalize the measure of the entire set of agents to be equal to one and measure the fraction of agents in any subset by its Lebesgue measure. The assumption of a large number of individually insignificant agents ensures perfect competition, that is, no one has enough market power to affect the equilibrium terms of trade.

The agents' time preferences are subject to a random shock at the beginning of date 1. With probability λ an agent is an early consumer, who only values consumption at date 1; with probability $(1-\lambda)$ he is a late consumer who only values consumption at date 2. The agent's (random) utility function $u(c_1, c_2)$ is defined by

$$u(c_1, c_2) = \begin{cases} U(c_1) & \text{w.pr. } \lambda \\ U(c_2) & \text{w.pr. } 1-\lambda, \end{cases}$$

where $c_t \geq 0$ denotes consumption at date $t = 1, 2$ and $U(\cdot)$ is a neoclassical utility function (increasing, strictly concave, twice continuously differentiable). Because there is a large number of agents and their preference shocks are independent, the 'law of large numbers' holds and we assume that the fraction of early consumers is constant and equal to the probability of being an early consumer. Then, although each agent is uncertain about his type, early or late, there is no uncertainty about the proportion of each type in the population. With probability one there is a fraction λ of early consumers and a fraction $1-\lambda$ of late consumers.

This simple example illustrates the problem of liquidity preference. If an agent knew his type at date 0, he could invest in the asset that would provide him with consumption exactly when he needed it. For example, if he were an early consumer, he could invest his endowment in the short asset and it would produce a return of one unit at date 1, exactly when he wanted to consume it. If he were a late consumer, he could invest his endowment in the long asset, which yields a higher rate of return, and it would produce R units at date 2 exactly when he wanted to consume it. The problem is precisely that he does not know his preferences regarding the timing of consumption when he has to make the investment decision. If he invests in the short asset and turns out to be a late consumer, he will regret not having invested in the higher-yielding long asset. If he invests in the long asset and turns out to be an early consumer, he will have nothing to consume and will clearly regret not having invested in the short asset. Even if he invests in a mixture of the two assets, he will still have some regrets with positive probability. It is this problem – the mismatch between asset maturity and time preferences – that financial intermediation is designed to solve.

3.2 MARKET EQUILIBRIUM

Before we consider the use of intermediaries to solve the liquidity problem, we consider a market solution. When we laid out the problem of matching investment maturities and time preferences in Chapter 2, we assumed that the agent existed in a state of autarky, that is, he was unable to trade assets and had to consume the returns generated by his own portfolio. The assumption of autarky is unrealistic. An agent often has the option of selling assets in order to realize their value in a liquid form. In fact, one of the main purposes of asset markets is to provide liquidity to agents who may be holding otherwise illiquid assets. So it is interesting to consider what would happen if long-term assets could be sold (liquidated) on markets and see whether this solves the problem of matching maturities to time preferences.

In this section, we assume that there exists a market on which an agent can sell his holding of the long asset at date 1 after he discovers his true type. Then, if he discovers that he is an early consumer, he can sell his holding of the long asset at the prevailing price and consume the proceeds. The existence of an asset market transforms the illiquid long asset into a liquid asset, in the sense that it can be sold for a predictable and sure price if necessary.

The possibility of selling the long asset at date 1 provides some insurance against liquidity shocks. Certainly, the agent must be at least as well off as he is in autarky and he may be better off. However, the asset market at date 1 cannot do as well as a complete set of contingent claims markets. For example, an agent might want to insure himself by trading goods for delivery at date 1 contingent on the event that he is an early or late consumer. As we shall see, if an agent can trade this sort of contingent claims at date 0, before his type is known, he can synthesize an optimal risk contract. The absence of markets for contingent claims in the present set up explains the agent's failure to achieve the first best.

At date 0, an investor has an endowment of one unit of the good which can be invested in the short asset or the long asset to provide for future consumption. Suppose he invests in a portfolio (x, y) consisting of x units of the long asset and y units of the short asset. His budget constraint at date 0 is

$$x + y \leq 1.$$

At date 1 he discovers whether he is an early or late consumer. If he is an early consumer, he will liquidate his portfolio and consume the proceeds. The price of the long asset is denoted by P. The agent's holding of the short asset provides him with y units of the good and his holding of the long asset can be sold for Px units of the good. Then his consumption at date 1 is given by the budget constraint

$$c_1 = y + Px.$$

If the agent is a late consumer, he will want to rebalance his portfolio. In calculating the optimal consumption for a late consumer, there is no essential loss of generality in assuming that he always chooses to invest all of his wealth in the long asset at date 1. This is because the return on the short asset (between dates 1 and 2) is weakly dominated by the return on the long asset. To see this, note that if, contrary to our claim, the return on the short asset were greater than the return on the long asset, no one would be willing to hold the long asset at date 1 and the asset market could not clear. Thus, in equilibrium, either the two rates of return are equal or the long asset dominates the short asset (and no one holds the short asset at date 1). This ordering of rates of return implies that $P \leq R$ in equilibrium and that it is weakly optimal for the late consumer to hold only the long asset at date 1. Then the quantity of the long asset held by the late consumer is $x + y/P$ since he initially held x units of the long asset and he exchanges the return on the short asset for y/P units of the long asset. His consumption at date 2 will be given by the budget constraint

$$c_2 = \left(x + \frac{y}{P}\right) R.$$

From the point of view of an investor at date 0, the objective is to choose a portfolio (x, y) to maximize the investor's expected utility

$$\lambda U\left(y + Px\right) + (1 - \lambda)\, U\left[\left(x + \frac{y}{P}\right) R\right].$$

The investor's decision at date 0 can be simplified if we note that, in equilibrium, the price of the long asset must be $P = 1$. To see this, note that if $P > 1$ then the long asset dominates the short asset and no one will hold the short asset at date 0. In that case, early consumers will be offering the long asset for sale but there will be no buyers at date 1. Then the price must fall to $P = 0$, a contradiction. On the other hand, if $P < 1$, then the short asset dominates the long asset and no one will hold the long asset at date 0. Early consumers will consume the returns on the short asset at date 1, realizing a return of $1 > P$; late consumers will try to buy the long asset to earn a return of $R/P > R$. Since no one has any of the long asset, the price will be bid up to $P = R$, another contradiction. Thus, $P = 1$ in equilibrium. At this price, the two assets have the same returns and are perfect substitutes. The agent's portfolio choice becomes immaterial. The agent's consumption is

$$c_1 = x + Py = x + y = 1$$

at date 1 and

$$c_2 = \left(x + \frac{y}{P}\right) R = (x + y)R = R$$

at date 2. Then the equilibrium expected utility is

$$\lambda U(1) + (1 - \lambda)U(R).$$

This level of welfare serves as a benchmark against which to measure the value of having a banking system that can provide insurance against liquidity (preference) shocks.

The value of the market

It is obvious that the investor is at least as well off when he has access to the asset market as he is in autarky. Typically, he will be better off. To show this, we have to compare the market equilibrium allocation $(c_1, c_2) = (1, R)$ with the set of allocations that are feasible in autarky.

As we saw in Chapter 2, if the consumer does not have access to the asset market and is forced to remain in autarky, his consumption as an early consumer would be equal to his investment y in the safe asset, and his consumption as a late consumer would be equal to the return $Rx = R(1 - y)$ on his investment in the long asset plus his investment in the safe asset y which can be re-invested at the second date. Thus, the feasible consumption bundles have the form

$$c_1 = y$$
$$c_2 = y + R(1 - y)$$

for some feasible value of y between 0 and 1. The set of such consumption bundles is illustrated in Figure 3.1.

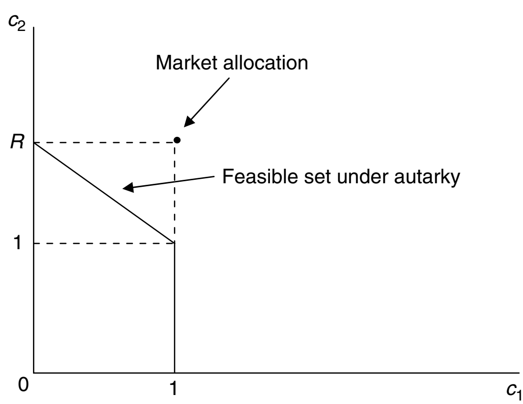

Figure 3.1. Comparison of the market allocation with the feasible set under autarky.

As the figure illustrates, the maximum value of early consumption is attained when $y = 1$ and $c_1 = 1$ and the maximum value of late consumption is attained when $y = 0$ and $c_2 = R$. The market allocation $(c_1, c_2) = (1, R)$ dominates every feasible autarkic allocation, that is, it gives strictly greater consumption at one date and typically gives greater consumption at both. Thus, access to the asset market does increase expected utility.

3.3 THE EFFICIENT SOLUTION

The market provides liquidity by allowing the investor to convert his holding of the long asset into consumption at the price $P = 1$ at the second date. Because the asset market is perfectly competitive, the investor can buy and sell any amount of the asset at the equilibrium price. This means that the market is **perfectly liquid** in the sense that the price is insensitive to the quantity of the asset that is traded. However, it turns out that the provision of liquidity is inefficient. We will discuss the reasons for this in greater detail later, but the short explanation for this inefficiency is that the set of markets in the economy described above is **incomplete**. In particular, there is no market at date 0 in which an investor can purchase **the good for delivery at date** 1 **contingent on his type**. If such a market existed, the equilibrium would be quite different.

We take as our definition of the efficient provision of liquidity, the level of welfare that could be achieved by a central planner who had command of all the allocation decisions in the economy. To begin with, we assume that the planner has complete information about the economy, including the ability to tell who is an early consumer and who is a late consumer. This important assumption will later be relaxed.

The central planner chooses the amount per capita x invested in the long asset and the amount per capita y invested in the short asset. Then he chooses the consumption per capita c_1 of the early consumers at date 1 and the consumption per capita c_2 of the late consumers at date 2. The central planner is not bound to satisfy any equilibrium conditions. He is only constrained by the condition that the allocation he chooses must be feasible. At date 0 the feasibility condition is simply that the total invested per capita must be equal to the per capita endowment:

$$x + y = 1. \tag{3.1}$$

At the second date, the feasibility condition is that the total consumption per capita must be less than or equal to the return on the short asset. Since the fraction of early consumers is λ and each one is promised c_1 the consumption per capita (i.e. per head of the entire population) is λc_1. Then the feasibility condition is

$$\lambda c_1 \leq y. \tag{3.2}$$

If this inequality is strict, some of the good can be re-invested in the short asset and consumed at date 2. Thus, the total available at date 2 (expressed in per capita terms) is $Rx + (y - \lambda c_1)$. The total consumption per capita (i.e. per head of the entire population) at date 2 is $(1 - \lambda)c_2$ since the fraction of late consumers is $1 - \lambda$ and each of them is promised c_2. So the feasibility condition is

$$(1 - \lambda)c_2 \leq Rx + (y - \lambda c_1),$$

which can also be written as

$$\lambda c_1 + (1 - \lambda)c_2 \leq Rx + y. \tag{3.3}$$

The planner's objective is to choose the investment portfolio (x, y) and the consumption allocation (c_1, c_2) to maximize the typical investor's expected utility

$$\lambda U(c_1) + (1 - \lambda)\, U(c_2),$$

subject to the various feasibility conditions (3.1)–(3.3).

This looks like a moderately complicated problem, but it can be simplified if we use a little common sense. The first thing to note is that it will never be optimal to carry over any of the short asset from date 1 to date 2. To see this, suppose that $y > \lambda c_1$ so that some of the good is left over at date 1. Then we could reduce the amount invested in the short asset at date 0 by $\varepsilon > 0$ say and invest it in the long asset instead. At date 2, there would be ε less of the short asset but ε more of the long asset. The net change in the amount of goods available would be $R\varepsilon - \varepsilon = (R - 1)\varepsilon > 0$, so it would be possible to increase the consumption of the late consumers without affecting the consumption of the early consumers. This cannot happen in an optimal plan, so it follows that in any optimal plan we must have $\lambda c_1 = y$ and $(1 - \lambda)c_2 = Rx$. Thus, once x and y are determined, the optimal consumption allocation is also determined by the relations

$$c_1 = \frac{y}{\lambda};$$

$$c_2 = \frac{Rx}{1 - \lambda}.$$

(Recall that y is the return on the short asset *per head of the entire population*, whereas c_1 is the consumption of a typical early consumer, so c_1 is greater than y. Similarly, the consumption c_2 of a typical late consumer is greater than the return on the long asset *per head of the entire population.*) If we substitute these expressions for consumption into the objective function and use the date-0 feasibility condition to write $x = 1 - y$, we can see that the planner's problem boils down to choosing y in the interval [0, 1] to maximize

$$\lambda U\left(\frac{y}{\lambda}\right) + (1-\lambda)\, U\left(\frac{R(1-y)}{1-\lambda}\right). \tag{3.4}$$

Ignoring the possibility of a boundary solution, where $y = 0$ or $y = 1$, a necessary condition for an optimal choice of y is that the derivative of the function in (3.4) be zero. Differentiating this function and setting the derivative equal to zero yields

$$U'\left(\frac{y}{\lambda}\right) - U'\left(\frac{R(1-y)}{1-\lambda}\right) R = 0,$$

or, substituting in the consumption levels,

$$U'(c_1) = U'(c_2)\, R. \tag{3.5}$$

There are several interesting observations to be made about this first-order condition. First, the value of λ does not appear in this equation: λ drops out when we differentiate the objective function. The intuition for this result is that λ appears symmetrically in the objective function and in the feasibility conditions. An increase in λ means that early consumers get more weight in the objective function but it also means that they cost more per capita to feed. These two effects cancel out and leave the optimal level of consumption unchanged.

The inefficiency of the market solution

The second point to note concerns the (in)efficiency of the market solution. The set of feasible consumption allocations for the planner's problem is illustrated in Figure 3.2. For each choice of y in the interval between 0 and 1, the consumption allocation is defined by the equations

$$c_1 = \frac{y}{\lambda};$$
$$c_2 = \frac{R(1-y)}{(1-\lambda)}.$$

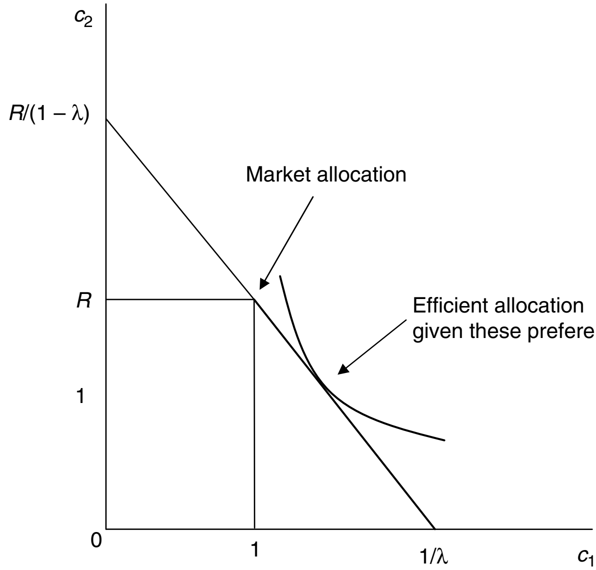

Figure 3.2. Comparison of the market allocation with the efficient allocation.

The consumption of the early consumers is maximized by putting $y = 1$, in which case $(c_1, c_2) = (1/\lambda, 0)$. Similarly, the consumption of the late consumers is maximized by putting $y = 0$, in which case $(c_1, c_2) = (0, R/(1-\lambda))$. Since the equations for consumption are linear in y, we can attain any point on the line segment joining $(1/\lambda, 0)$ and $(0, R/(1-\lambda))$. This feasible frontier is described in the figure. The efficient point is determined by the tangency of the consumers' indifference curve with the feasible frontier. Depending on the consumers' preferences, the point of tangency could occur anywhere along the feasible frontier.

The market allocation occurs on the feasible frontier. Simply put $y = \lambda$ and we get

$$(c_1, c_2) = \left(\frac{y}{\lambda}, \frac{R(1-y)}{(1-\lambda)}\right) = (1, R).$$

This allocation could be efficient but typically it will not be. To see this, suppose that by some chance the market equilibrium resulted in the same allocation as the planner's problem. Then the first-order condition (3.5) becomes

$$U'(1) = U'(R)R.$$

In some special cases this condition may be satisfied. For example, suppose that the investor has a logarithmic utility function $U(c) = \ln c$ so that $U'(c) = 1/c$. Substituting this expression in the preceding equation, the left hand side becomes $U'(1) = 1$ and the right hand side becomes $U'(R)R = (1/R)\,R = 1$. In this particular case, the market provision of liquidity is efficient: a central planner could do no better than the market. For other utility functions, this would not be the case. Suppose, for example, that the investor's utility function belongs to the constant relative risk aversion class

$$U(c) = \frac{1}{1-\sigma} c^{1-\sigma}$$

where $\sigma > 0$ is the degree of relative risk aversion. Then $U'(c) = c^{-\sigma}$ and substituting this into the necessary condition for efficiency, we see that the left hand side becomes $U'(1) = 1$ and the right hand side becomes $U'(R)R = R^{-\sigma}R = R^{1-\sigma}$. Except in the case $\sigma = 1$ (which corresponds to the logarithmic case), $R^{1-\sigma} \neq 1$ and the allocation chosen by the planner must be different from the market allocation. So for any degree of risk aversion different from 1 the planner achieves a strictly better level of expected utility than the market.

Liquidity insurance

A third insight that can be derived from the first-order condition (3.5) concerns the provision of insurance against liquidity shocks, that is, the event of being an early consumer. Even in the efficient allocation the individual faces uncertainty about consumption. The first-order condition $U'(c_1) = RU'(c_2)$ implies that $c_1 < c_2$ so the individual's consumption will be higher or lower depending on whether he is a late or an early consumer.

In the market allocation, the early consumers get 1 and the late consumers get $R > 1$ so it is clearly worse to be an early consumer than a late consumer. A risk averse investor would like to have more consumption as an early consumer and less as a late consumer, *assuming that the expected value of his consumption remains the same.* An interesting question is whether the planner provides insurance against the liquidity shock by reducing the volatility of consumption, that is, increasing consumption of early consumers and reducing consumption of late consumers *relative to the market solution.* In Figure 3.2, the consumption allocations that lie to the right of the market allocation all satisfy $c_1 > 1$ and $c_2 < R$, that is, they have less uncertainty about consumption than the market allocation. On the other hand, they also have lower expected consumption.

It is interesting to see under what conditions the optimal deposit contract gives the consumer more consumption at date 1 than the market solution, that is, when are the higher returns from the long asset shared between the early and late consumers. Substituting from the budget constraints into the first-order condition we get the equation

$$U'\left(\frac{y}{\lambda}\right) = RU'\left(\frac{(1-y)}{(1-\lambda)}R\right).$$

If $y = \lambda$ this condition reduces to $U'(1) = RU'(R)$, so a necessary and sufficient condition for $c_1 = y/\lambda > 1$ and $c_2 = R(1-y)/(1-\lambda) < R$ is that

$$U'(1) > RU'(R).$$

A sufficient condition is that $cU'(c)$ be decreasing in c, which is equivalent to saying that the relative risk aversion is greater than one:

$$\eta(c) \equiv -\frac{cU''(c)}{U'(c)} > 1.$$

If this inequality is reversed and $\eta(c) < 1$, early consumers get less and late consumers get more than in the benchmark case, that is,

$$c_1 = y/\lambda < 1,$$
$$c_2 = R(1-y)/(1-\lambda) > R.$$

These results have an intuitive explanation. The logarithmic utility function, which has a degree of relative risk aversion equal to unity, marks the boundary between two different cases. If relative risk aversion equals one, as in the natural log case, the market outcome is efficient and, in particular, the market's provision of liquidity is optimal. If relative risk aversion is greater than one, the market provision of liquidity is inefficient. An efficient allocation should provide more insurance by increasing the consumption of the early consumers and reducing the consumption of the late consumers. If relative risk aversion is less than one, there is paradoxically too much liquidity in the sense that efficiency requires a reduction in the consumption of the early consumers and an increase in the consumption of the late consumers.

This last result alerts us to the fact that insurance is costly. In order to provide more consumption to the early consumers, it is necessary to hold more of the short asset and, hence, less of the long asset. Since the long asset's return is greater than the short asset's, the increase in the amount of the short asset in the planner's portfolio must reduce average consumption across the two dates. As long as relative risk aversion is greater than one, the benefit from insurance is greater than the cost, at least to start with. If the relative risk aversion is less than one, the benefits of greater insurance are not worth the cost and, indeed, the efficient allocation requires the planner to increase the risk borne by the investors in order to capture the increased returns from holding the long asset.

Why do the investors hold the wrong portfolio in the market solution? Their portfolio decision is dependent upon the market price for the long asset $P = 1$. This price is determined by the condition that investors at date 0 must be willing to hold both assets. The market does not reveal how much investors would be willing to pay for the asset contingent on knowing their type. Consequently, the price does not reflect the value of being able to sell the long asset as an early consumer or being able to buy it as a late consumer.

Complete markets

We mentioned earlier (in Section 3.2) the possibility of introducing markets that would allow individual agents to trade at date 0 claims on date-1 consumption contingent on their type, early or late. The existence of such markets would achieve the same allocation of risk and the same portfolio investment as the central planner. Unlike the model economy in which there are no contingent markets, an economy with markets for individual contingencies would signal the correct value of each asset to the market, in particular the value of liquidity, and ensure that the efficient allocation was achieved. To see this, suppose that an individual can purchase date-1 consumption at a price q_1 if he is early and q_2 if he is late. Note that these prices are measured in terms of the good at date 0. The implicit price of goods at date 2 in terms of goods at date 1 is $p = P/R$, as usual. Then the budget constraint for an individual at date 0 is

$$q_1 \lambda C_1 + q_2 p (1 - \lambda) C_2 \leq 1. \tag{3.6}$$

The left hand side represents the present value (at date 0) of expected consumption (since there is no aggregate uncertainty, each individual only pays for the expected value of his demand for goods at each date). With probability λ he demands C_1 units of date-1 consumption and the present value of λC_1 is $q_1 \lambda C_1$. Similarly, with probability $1 - \lambda$ he demands C_2 units of date-2 consumption and the present value of $(1 - \lambda) C_2$ is $q_2 p (1 - \lambda) C_2$.

The individual chooses (C_1, C_2) to maximize $\lambda U (C_1) + (1 - \lambda) U (C_2)$ subject to (3.6) and the solution must satisfy the first-order conditions

$$\lambda U' (C_1) = \mu q_1 \lambda$$

and

$$(1 - \lambda) U' (C_2) = \mu q_2 p (1 - \lambda)$$

where $\mu > 0$ is the Lagrange multiplier associated with the constraint. Then

$$\frac{U'(C_1)}{U'(C_2)} = \frac{q_1}{q_2 p}.$$

Since the investment technology exhibits constant returns to scale, the equilibrium prices must satisfy two no-arbitrage conditions. To provide one unit of the good at date 1, it is necessary to invest 1 unit in the short asset at date 0. Thus, there are zero profits from investing in the short asset if and only if $q_1 = 1$. Similarly, to provide one unit of the good at date 2, it is necessary to invest $1/R$ units in the long asset at date 0. Thus, there are zero profits from investing in the long asset if and only if $pq_2 = 1/R$. This implies that

$$\frac{U'(C_1)}{U'(C_2)} = R,$$

the condition required for efficient risk sharing.

Private information and incentive compatibility

So far we have assumed that the central planner knows everything, including whether an investor is an early or late consumer. This allows the planner to assign different levels of consumption to early and late consumers. Since an investor's time preferences are likely to be private information, the assumption that the planner knows the investor's type is restrictive. If we relax this assumption, we run into a problem: if time preferences are private information, how can the planner find out who is an early consumer and who is a late consumer? The planner can rely on the individual truthfully revealing his type if and only if the individual has no incentive to lie. In other words, the allocation chosen by the planner must be **incentive-compatible.**

In the present case, it is quite easy to show that the optimal allocation is incentive compatible. First, the early consumers have no opportunity to misrepresent themselves as late consumers. A late consumer is given c_2 at date 2 and since the early consumer only values consumption at date 1 he would certainly be worse off if he waited until date 2 to receive c_2. The late consumer poses more of a problem. He could pretend to be an early consumer, receive c_1 at date 1 and store it until date 2 using the short asset. To avoid giving the late consumer an incentive to misrepresent his preferences, he must receive at least as much consumption as the early consumer. This means that the allocation is incentive compatible if and only if

$$c_1 \leq c_2. \qquad (3.7)$$

Fortunately for us, if we consult the first-order condition for the planner's problem, equation (3.5), we find that it implies that $c_1 < c_2$ since $R > 1$ and $U''(c) < 0$ so that the optimal allocation is automatically incentive-compatible.

The allocation that optimizes the typical investor's welfare subject to the incentive constraint (3.7) is called **incentive-efficient**. In this case, because the incentive constraint is not binding at the optimum, the incentive-efficient allocation is the same as the first-best allocation.

Although the incentive-compatibility condition does not have any effect on the optimal risk-sharing allocation, private information plays an important role in the account of banking that we give in the sequel. In particular, the fact that a bank cannot distinguish early and late consumers means that all consumers can withdraw from the bank at date 1 and this is a crucial feature of the model of bank runs.

3.4 THE BANKING SOLUTION

A bank, by pooling the depositors' investments, can provide insurance against the preference shock and allow early consumers to share the higher returns of the long asset. The bank takes one unit of the good from each agent at date 0 and invests it in a portfolio (x, y) consisting of x units of the long asset and y units of the short asset. Because there is no aggregate uncertainty, the bank can offer each consumer a non-stochastic consumption profile (c_1, c_2), where c_1 is the consumption of an early consumer and c_2 is the consumption of a late consumer. We can interpret (c_1, c_2) as a deposit contract under which the depositor has the right to withdraw either c_1 at date 1 or c_2 at date 2, but not both.

There is assumed to be free entry into the banking sector. Competition among the banks forces them to maximize the ex ante expected utility of the typical depositor subject to a zero-profit (feasibility) constraint. In fact, the bank is in exactly the same position as the central planner discussed in the previous section. At date 0 the bank faces a budget constraint

$$x + y \leq 1. \tag{3.8}$$

At date 1, the bank faces a budget constraint

$$\lambda c_1 \leq y. \tag{3.9}$$

Recalling that it is never optimal to carry consumption over from date 1 to date 2 by holding the short asset, we can write the budget constraint for the bank at the third date as

$$(1 - \lambda)\, c_2 \leq Rx. \tag{3.10}$$

Formally, the bank's problem is to maximize the expected utility of the typical depositor

$$\lambda U(c_1) + (1 - \lambda) U(c_2)$$

subject to the budget constraints (3.8)–(3.10).

We do not explicitly impose the incentive-compatibility constraint because, as we saw previously, the solution to the unconstrained optimization problem will automatically satisfy the incentive constraint

$$c_1 \leq c_2.$$

So the bank is able to achieve the first-best allocation on behalf of its customers.

It is worth pausing to note how this account of bank behavior implements three of the four elements of banking theory mentioned at the beginning of this chapter.

- It provides a model of the **maturity structure** of bank assets, in which less liquid assets earn higher returns. In this case, there are two bank assets, the liquid short asset, which yields a return of 1, and the illiquid long asset, which yields a return of $R > 1$.
- It provides a theory of **liquidity preference**, modeled as uncertainty about the timing of consumption. The maturity mismatch arises because an investor is uncertain of his preferences over the timing of consumption at the moment when an investment decision has to be made.
- It represents the bank as an intermediary that provides **insurance** to depositors against liquidity (preference) shocks. By pooling his resources with the bank's and accepting an insurance contract in the form of promises of consumption contingent on the date of withdrawal, the investor is able to achieve a better combination of liquidity services and returns on investment than he could achieve in autarky or in the asset market.

The properties of the efficient allocation, derived in the preceding section, of course apply to the banking allocation, so we will not repeat them here. Instead, we want to focus on the peculiar fragility of the arrangement that the bank has instituted in order to achieve optimal risk sharing.

3.5 BANK RUNS

At the beginning of this chapter, we mentioned four contributions to banking theory made by the seminal papers of Bryant (1980) and Diamond and Dybvig (1983). We have discussed the first three and now we turn to the fourth, namely, the explanation of bank runs. In this section, we develop a model of bank runs as panics or self-fulfilling prophecies. Later we shall consider bank runs as the result of fundamental forces arising in the course of the business cycle.

Suppose that (c_1, c_2) is the optimal deposit contract and (x, y) is the optimal portfolio for the bank. In the absence of aggregate uncertainty, the portfolio (x, y) provides just the right amount of liquidity at each date assuming that the early consumers are the only ones to withdraw at date 1 and the late consumers all withdraw at date 2. This is an equilibrium in the sense that the bank is maximizing its objective, the welfare of the typical depositor, and the early and late consumers are timing their withdrawals to maximize their consumption.

So far, we have treated the long asset as completely illiquid: there is no way that it can be converted into consumption at date 1. Suppose, instead, that there exists a **liquidation technology** that allows the long-term investment to be terminated prematurely at date 1. More precisely, we assume that

- if the long asset is liquidated prematurely at date 1, one unit of the long asset yields $r \leq 1$ units of the good.

Under the assumption that the long asset can be prematurely liquidated, with a loss of $R - r$ per unit, there exists another equilibrium if we also assume that the bank is required to liquidate whatever assets it has in order to meet the demands of the consumers who withdraw at date 1. To see this, suppose that all depositors, whether they are early or late consumers, decide to withdraw at date 1. The liquidated value of the bank's assets at date 1 is

$$rx + y \leq x + y = 1$$

so the bank cannot possibly pay all of its depositors more than one unit at date 1. In the event that $c_1 > rx + y$, the bank is insolvent and will be able to pay only a fraction of the promised amount. More importantly, all the bank's assets will be used up at date 1 in the attempt to meet the demands of the early withdrawers. Anyone who waits until the last period will get nothing. Thus, given that a late consumer thinks everyone else will withdraw at date 1 it is optimal for a late consumer to withdraw at date 1 and save the proceeds until date 2. Thus, bank runs are an equilibrium phenomenon. The following payoff matrix illustrates the two equilibria of this coordination game. The rows correspond to the decision of an individual late consumer and the columns to the decision of all the other late consumers. (Note: this is **not** a 2 × 2 game; the choice of column represents the actions of all but one late consumer.) The ordered pairs are the payoffs for the distinguished late consumer (the first element) and the typical late consumer (the second element).

	Run	No Run
Run	$(rx + y, rx + y)$	(c_1, c_2)
No Run	$(0, rx + y)$	(c_2, c_2)

It is clear that if

$$0 < rx + y < c_1 < c_2$$

then (Run, Run) is an equilibrium and (No Run, No Run) is also an equilibrium.

The preceding analysis (of a bank run) is predicated on the assumption that the bank liquidates all of its assets in order to meet the demand for liquidity at date 1. This may be the result of legal restrictions. For example, bankruptcy law or regulations imposed by the banking authority may require that if any claim is not met, the bank must wind up its business and distribute the liquidated value of its assets to its creditors. Some critics of the Diamond–Dybvig model have argued that bank runs can be prevented by suspension of convertibility. If banks commit to suspend convertibility (i.e. they refuse to allow depositors to withdraw), once the proportion of withdrawals is equal to the proportion of early consumers, then late consumers will not have an incentive to withdraw. A late consumer knows that the bank will not have to liquidate the long asset and will have enough funds to pay him the higher promised amount in the second period. If such an agent were to join the run on the bank in the middle period, he would be strictly worse off than if he waited to withdraw at the last date.

To answer the criticism that suspension of convertibility solves the bank-run problem, Diamond and Dybvig (1983) proposed a *sequential service constraint.* Under this assumption, depositors reach the bank's teller one after another and withdraw c_1 until the bank is unable to meet any further demand. The sequential service constraint has two effects. It forces the bank to deplete its resources and it gives depositors an incentive to run early in hopes of being at the front of the queue. The bank cannot use suspension of convertibility to prevent runs since it does not find out a run is in progress until it is too late to stop it.

Another point to note about the suspension of convertibility is that it solves the bank-run problem only if the bank knows the proportion of early consumers. If the proportion of early consumers is random and the realization is not known by banks, the bank cannot in general implement the optimal allocation by using suspension of convertibility.

3.6 EQUILIBRIUM BANK RUNS

The analysis offered by Diamond and Dybvig pinpoints the fragility of banking arrangements based on liquid liabilities and illiquid assets, but it does not provide a complete account of equilibrium in the banking sector. Instead, it assumes that the bank's portfolio (x, y) and deposit contract (c_1, c_2) are chosen at date 0 in the expectation that the first-best allocation *will* be achieved. In other words, the bank run at date 1, if it occurs, is entirely unexpected at date 0. Taking the decisions at date 0 as given, we can define an equilibrium at date 1 in which a bank run occurs; but this is not the same thing as showing that there exists an equilibrium *beginning at date* 0 in which a bank run is expected to occur. If banks anticipated the possibility of a bank run, their decisions at date 0 would be different and that in turn might affect the probability or even the possibility of a bank run at date 1. What we need is an equilibrium account of bank runs that describes consistent decisions at all three dates. In this section, we provide a coherent account of bank runs as part of an equilibrium that includes the decisions made at date 0. We proceed by establishing a number of facts or properties of equilibrium bank runs before describing the overall picture.

The impossibility of predicting bank runs

The first thing we need to notice in constructing an equilibrium account of bank runs is that a bank run cannot occur with probability one. If a bank run is certain at date 0, the bank knows that each unit of the good invested in the long asset will be worth r units at date 1. If $r < 1$, the long asset is dominated by the short asset and the bank will not invest in the long asset at all. If $r = 1$, the two assets are for all intents and purposes identical. In either case, the optimal deposit contract is $(c_1, c_2) = (1, 1)$ and there is no motive for a bank run: a late consumer will get the same consumption whether he joins the run or not. So, the best we can hope for is a bank run that occurs *with positive probability*.

The role of sunspots

The uncertainty of the bank run introduces a new element in our theory. In the current model, there is no uncertainty about aggregate fundamentals, such as asset returns, the proportion of early consumers, and so on. The kind of uncertainty we are contemplating here is *endogenous*, in the sense that it is not explained by shocks to the fundamentals of the model. How can we explain such uncertainty? Traditional accounts of bank runs often referred to "mob psychology." Modern accounts explain it as the result of coordination among individuals that is facilitated by extraneous variables called "sunspots." We shall have more to say about the distinction between these different types of uncertainty in Chapter 5. For the moment, it is enough to note that the uncertainty is not explained by exogenous shocks, but is completely consistent with the requirements of equilibrium, namely, that every individual is maximizing his expected utility and that markets clear.

We begin by hypothesizing that a bank run occurs at date 1 with probability $0 < \pi < 1$. To be more concrete, we can assume there is some random variable (sunspot) that takes two values, say, high and low, with probabilities π and $1 - \pi$, respectively. When the realization of the random variable is high, depositors run on the bank and when it is low, they do not. Note that the random variable has no direct effect on preferences or asset returns. It is merely a device for coordinating the decisions of the depositors. It is rational for the depositors to change their behavior depending on the value of the sunspot merely because they expect everyone else to do so.

The bank's behavior when bank runs are uncertain

The expectation of a bank run at date 1 changes the bank's behavior at date 0. As usual, the bank must choose a portfolio (x, y) and propose a deposit contract (c_1, c_2), but it does so in the expectation that the consumption stream (c_1, c_2) will be achieved only if the bank is solvent. In the event of a bank run, on the other hand, the typical depositor will receive the value of the liquidated portfolio $rx + y$ at date 1. This means that

- with probability π there is a bank run and the depositor's consumption is $rx + y$, regardless of his type;
- with probability $(1 - \pi)\lambda$ there is no run, the depositor is an early consumer, and his consumption at date 1 is c_1;
- and with probability $(1 - \pi)(1 - \lambda)$ there is no run, the depositor is a late consumer, and his consumption at date 2 is c_2.

The outcomes of the bank's decisions when runs are anticipated are illustrated in Figure 3.3.

The optimal portfolio

If c_1 denotes the payment to early consumers when the bank is solvent and (x, y) denotes the portfolio, then the expected utility of the representative

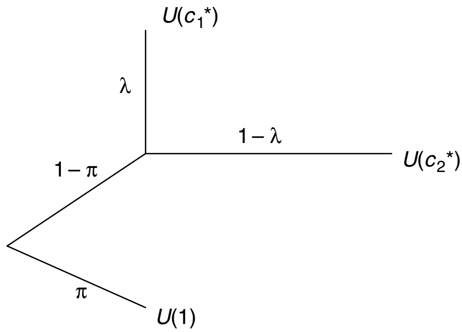

Figure 3.3. Equilibrium outcome when runs are anticipated with probability $\pi > 0$.

depositor can be written

$$\pi U(y + rx) + (1 - \pi) \{\lambda U(c_1) + (1 - \lambda) U(c_2)\}.$$

Now suppose that we increase y by a small amount $\varepsilon > 0$ and decrease x by the same amount. We increase λc_1 by ε and reduce $(1 - \lambda) c_2$ by $R\varepsilon$. This insures the feasibility constraints are satisfied at each date. Then the change in expected utility is

$$\pi U'(y + rx)(1 - r)\varepsilon + (1 - \pi)\left\{U'(c_1) - U'(c_2) R\right\}\varepsilon + o(\varepsilon).$$

The optimal portfolio must therefore satisfy the first-order condition

$$\pi U'(y + rx)(1 - r) + (1 - \pi) U'(c_1) = (1 - \pi) U'(c_2) R.$$

If $\pi = 0$ then this reduces to the familiar condition $U'(c_1) = U'(c_2) R$. These relations are graphed in Figure 3.4. The latter condition holds at y^* while the former holds at y^{**}. Thus, the possibility of a run increases the marginal value of an increase in y (the short asset has a higher return than the long asset in the bankruptcy state if $r < 1$) and hence increases the amount of the short asset held in the portfolio.

The optimal deposit contract

Our next task is to show that a bank run is possible when the deposit contract is chosen to solve the bank's decision problem. To maximize expected utility, the bank must choose the deposit contract $(c_1{}^*, c_2{}^*)$ to satisfy the first-order condition

$$U'(c_1{}^*) = RU'(c_2{}^*). \tag{3.11}$$

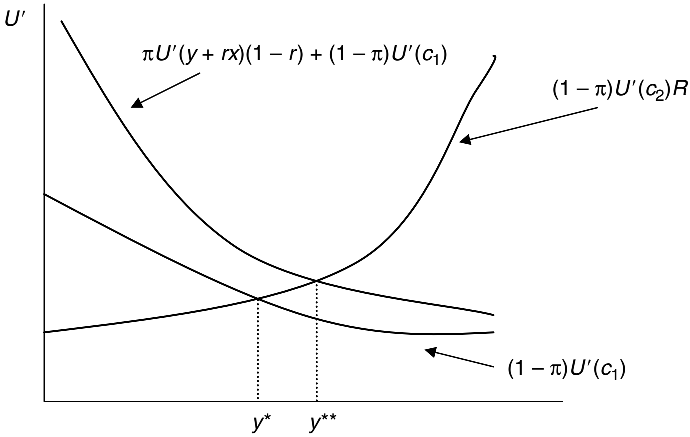

Figure 3.4. The determination of the optimal portfolio when bank runs are possible.

This condition, which is familiar from our characterization of the first best, plays a crucial role in determining whether the bank is susceptible to runs.

As we saw earlier, if relative risk aversion is greater than one, then the solution of the first-order condition (3.11) must satisfy the inequality

$$c_1 > 1.$$

This condition implies there exists the possibility of a run. If all the depositors try to withdraw at date 1, the total demand for consumption is $c_1^* > 1$ but the maximum that can be provided by liquidating all of the long asset is 1. However, there will be nothing left at date 2 so the depositors are better off joining the run than waiting until date 2 to withdraw.

In what follows, we assume that the agent's preferences satisfy the condition that

- relative risk aversion is greater than one, that is,

$$-\frac{U''(c)c}{U'(c)} > 1, \ \forall c > 0.$$

To simplify the characterization of the equilibrium, we only consider the special case in which the long asset, when liquidated prematurely, yields as much as the short asset. In other words,

- the liquidation value of the long asset is $r = 1$.

This implies that the long asset dominates the short asset so, without essential loss of generality, we can assume in what follows that the entire bank portfolio is invested in the long asset.

In the event of a bank run, the liquidated value of the bank's portfolio is one unit of the good, so every depositor's consumption is also one unit of the good. If the bank is solvent, the depositors receive the promised consumption profile (c_1, c_2). Since these quantities only apply in the event that the bank is solvent, they are chosen to maximize the typical consumer's expected utility in the event that the bank is solvent. The deposit contract must solve the decision problem

$$\begin{array}{ll} \max & \lambda U(c_1) + (1-\lambda) U(c_2) \\ \text{s.t.} & R\lambda c_1 + (1-\lambda)\, c_2 \leq R. \end{array}$$

To see why the budget constraint takes this form, note that the bank has promised a total of λc_1 units to the early consumers and this requires the bank to liquidate λc_1 units of the long asset at date 1. The amount of the long asset left is $(1 - \lambda c_1)$ and this produces $R\,(1 - \lambda c_1)$ units of consumption at date 2. Thus, the maximum amount that can be promised to late consumers $(1-\lambda)\, c_2$ must be less than or equal to $R\,(1 - \lambda c_1)$. In effect, one unit of consumption at date 1 is worth R units of consumption at date 2.

Equilibrium without runs

So far we have assumed that a run occurs with probability π and that the bank takes this possibility as given in choosing an optimal deposit contract; however, the bank can avoid a run by choosing a sufficiently "safe" contract. Remember that our argument for the existence of a run equilibrium at date 1 was based on the assumption that $c_1 > 1$. Thus, if all the late consumers join the run on the bank at date 1 there is no way that the bank can provide everyone with c_1. In fact, the bank will have to liquidate all its assets and even then can only give each withdrawer 1, the liquidated value of its portfolio. More importantly, since the bank's assets are exhausted at date 1, anyone waiting until date 2 to withdraw will receive nothing.

In order to remove this incentive to join the run, the bank must choose a deposit contract that satisfies the additional constraint $c_1 \leq 1$. If we solve the problem

$$\begin{array}{ll} \max & \lambda U(c_1) + (1-\lambda) U(c_2) \\ \text{s.t.} & R\lambda c_1 + (1-\lambda)\, c_2 \leq R \\ & c_1 \leq 1 \end{array}$$

we find the solution $\left(c_1^{**}, c_2^{**}\right) = (1, R)$. In this case, the bank will be able to give everyone the promised payment c_1 at date 1 and if any late consumers wait until date 2 to withdraw there will be enough left over to pay them at least $R > 1$. More precisely, if $1 - \varepsilon$ of the depositors withdraw at date 1, the bank has to liquidate $1 - \varepsilon$ units of the long asset, leaving ε units of the long asset to pay to the remaining late consumers. Then each consumer who withdraws at date 2 will receive $\varepsilon R/\varepsilon = R > 1$.

A characterization of regimes with and without runs

If the bank anticipates a run with probability π, then with probability π the depositor's consumption is 1, regardless of his type. With probability $1 - \pi$ there is no run and with probability λ the depositor is an early consumer and his consumption is c_1^* and with probability $1 - \lambda$ he is a late consumer and his consumption is c_2^*. The possible outcomes are illustrated in Figure 3.3 (above). The expected utility of the typical depositor will be

$$\pi U(1) + (1 - \pi) \left\{\lambda U(c_1^*) + (1 - \lambda)\, U\left(c_2^*\right)\right\},$$

and we have shown that the bank's choice of portfolio $\left(x, y\right) = (1, 0)$ and deposit contract $\left(c_1^*, c_2^*\right)$ will maximize this objective, taking the probability π of a run as given.

Alternatively, if the bank chooses a deposit contract that avoids all runs, the expected utility of the typical depositor is

$$\lambda U(c_1^{**}) + (1 - \lambda)\, U(c_1^{**}) = \lambda U(1) + (1 - \lambda)\, U(R).$$

Whether it is better for the bank to avoid runs or accept the risk of a run with probability π depends on a comparison of the expected utilities in each case. Precisely, it will be better to avoid runs if

$$\pi U(1) + (1 - \pi) \left\{\lambda U(c_1^*) + (1 - \lambda)\, U\left(c_2^*\right)\right\} > \lambda U(1) + (1 - \lambda)\, U(R).$$

Notice that the left hand side is a convex combination of the depositors' utility $U(1)$ when the bank defaults and their expected utility $\lambda U(c_1^*) + (1 - \lambda)\, U\left(c_2^*\right)$ when the bank is solvent. Now, the expected utility from the safe strategy $\lambda U(1) + (1 - \lambda)\, U(R)$ lies between these two values:

$$\begin{aligned} U(1) &< \lambda U(1) + (1 - \lambda)\, U(R) \\ &< \lambda U(c_1^*) + (1 - \lambda)\, U\left(c_2^*\right). \end{aligned}$$

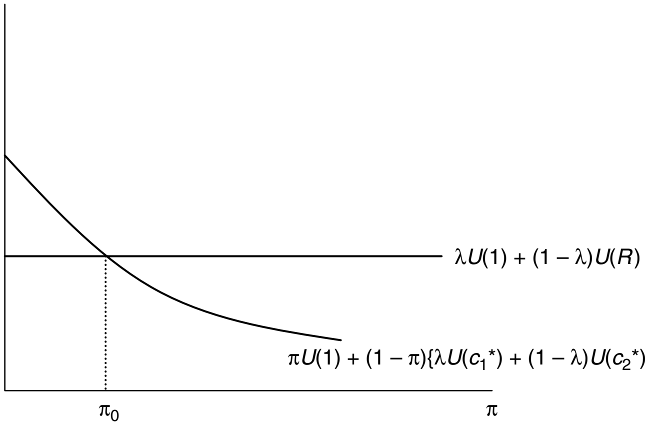

Figure 3.5. Determination of the regions of π supporting (respectively not supporting) runs.

So there exists a unique value $0 < \pi_0 < 1$ such that

$$\pi_0 U(1) + (1 - \pi_0) \left\{\lambda U(c_1^*) + (1 - \lambda) U\left(c_2^*\right)\right\} = \lambda U(1) + (1 - \lambda) U(R)$$

and the bank will be indifferent between the two strategies if $\pi = \pi_0$. Obviously, the bank will prefer runs if $\pi < \pi_0$ and will prefer no runs if $\pi > \pi_0$. These two regions are illustrated in Figure 3.5.

We have shown that as long as the probability of a bank run is sufficiently small, there will exist an equilibrium in which the bank is willing to risk a run because the cost of avoiding the run outweighs the benefit. In that case, there will be a run if the sunspot takes the high value and not otherwise. There is an upper bound (less than one) to the probability of a run, however. If the probability of a run is too high, the bank will take action to discourage a run and the depositors will find it optimal to withdraw at the correct date.

Note that we have not specified what the sunspot is. It could be any publicly observed random variable that takes on a particular value with probability $\pi < \pi_0$. If such a variable exists, then depositors can in principle coordinate on this variable to support an equilibrium bank run.

3.7 THE BUSINESS CYCLE VIEW OF BANK RUNS

The previous sections have outlined a Diamond–Dybvig style account of bank runs in which extrinsic uncertainty plays a crucial role. Runs occur in this framework because of late consumers' beliefs. If all the late consumers believe there will be a run, they will all withdraw their money in the middle period. If they do not believe a run will occur, they will wait until the last period to withdraw. In both cases, beliefs are self-fulfilling. In the last section we used the terminology of sunspots to explain how coordination occurs. Traditional accounts of bank runs often referred to "mob psychology" as the motive for the bank run or "panic." This view of bank runs as panics has a long history but it is not the only view. An alternative view of bank runs is that they are a natural outgrowth of weak fundamentals arising in the course of the *business cycle*. An economic downturn will reduce the value of bank assets, raising the possibility that banks will be unable to meet their commitments in the future. 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, crises are not random events, but a rational response to unfolding economic circumstances. In other words, they are an integral part of the business cycle.

In Chapter 1 we briefly discussed financial crises in the US during the National Banking Era from 1865 to 1914. Gorton (1988) conducted an empirical study to differentiate between the sunspot view and the business cycle view of banking crises using data from this period. He found evidence consistent with the view that banking panics are predicted by the business cycle. It is difficult to reconcile this finding with the notion of crises as "random" events. Table 3.1 shows the recessions and crises that occurred in the US during the National Banking Era. It also shows the corresponding percentage changes in the currency/deposit ratio and the change in aggregate GDP, as proxied by the change in pig iron production during these periods. The five worst recessions, as measured by the change in pig iron production, were accompanied by crises. In all, crises occurred in seven of the eleven cycles. Using the liabilities of failed businesses as a leading economic indicator, Gorton finds that crises were predictable events: whenever this leading economic indicator reached a certain threshold, a panic ensued. The stylized facts uncovered by Gorton thus suggest that, at least during the US National Banking Era, banking crises were intimately related to the business cycle rather than some extraneous random variable. Calomiris and Gorton (1991) consider a broad range of evidence from this period and conclude that the data do not support the "sunspot" view that banking panics are random events. Among other things, they find that for the five episodes they focus on, stock prices fell by the largest amount by far during the pre-panic periods.

In this section, we adapt our model to allow us to consider the fundamental or business cycle view of banking crises. In particular, instead of assuming the

Table 3.1. National Banking Era panics.

NBER Cycle Peak–Trough	Panic date	%Δ(Currency/ deposit)*	%Δ Pig iron†
Oct. 1873–Mar. 1879	Sep. 1873	14.53	−51.0
Mar. 1882–May 1885	Jun. 1884	8.80	−14.0
Mar. 1887–Apr. 1888	No Panic	3.00	−9.0
Jul. 1890–May 1891	Nov. 1890	9.00	−34.0
Jan. 1893–Jun. 1894	May 1893	16.00	−29.0
Dec. 1895–Jun. 1897	Oct. 1896	14.30	−4.0
Jun. 1899–Dec. 1900	No Panic	2.78	−6.7
Sep. 1902–Aug. 1904	No Panic	−4.13	−8.7
May 1907–Jun. 1908	Oct. 1907	11.45	−46.5
Jan. 1910–Jan. 1912	No Panic	−2.64	−21.7
Jan. 1913–Dec. 1914	Aug. 1914	10.39	−47.1

∗ Percentage change of ratio at panic date to previous year's average.
† Measured from peak to trough.
(Adapted from Table 1, Gorton 1988, p. 233).

long asset has a certain return, we assume that the return is risky. Here we are following the approach developed in Allen and Gale (1998) (cf. also Bryant 1980).

- The *long asset* is a constant returns to scale technology that takes one unit of the good at date 0 and transforms it into R_H units of the good at date 2 with probability π_H and R_L units with probability π_L. If the long asset is prematurely liquidated, one unit of the asset yields r units of the good at date 1. We assume that

$$R_H > R_L > r > 0.$$

An intermediary takes a deposit of one unit from the typical consumer and invests it in a portfolio consisting of y units of the safe, short asset and x units of the risky, long asset, subject to the budget constraint

$$x + y \leq 1.$$

In exchange, the intermediary offers the consumer a contract promising c_1 units of consumption if he withdraws at date 1 and c_2 units of consumption if he withdraws at date 2. As before we assume that the intermediary cannot observe the consumer's type (i.e. early or late) and so cannot make the contract contingent on that. A more stringent requirement is that the intermediary cannot make the deposit contract contingent on the state of nature or, equivalently, the return to the risky asset.

Free entry and competition among the intermediaries leads them to maximize the expected utility of their customers. This implies that the intermediaries will receive zero profits in equilibrium. In particular, this requires that the consumers receive the entire value of the remaining assets at date 2. Because the terminal value of the assets is uncertain, the intermediary will promise a large amount that will certainly exhaust the value of the assets at date 2. Without loss of generality we put $c_2 = \infty$ and, in what follows, we can characterize the deposit contract by the single parameter $c_1 = d$, where d stands for the face value of the deposit at date 1.

Introducing risk in the form of random asset returns does not eliminate the Diamond–Dybvig phenomenon of bank runs based on self-fulfilling expectations or coordination on sunspots. In fact, the Diamond–Dybvig model is just a special case of the current model with $R_H = R_L$. In order to distinguish this account of bank runs, we simply rule out the Diamond–Dybvig phenomenon by assumption and consider only **essential** bank runs, that is, runs that cannot be avoided. Loosely speaking, we assume that if there exists an equilibrium in which there is no bank run as well as one or more that have bank runs, then the equilibrium we observe is the one without a bank run rather than the one with a bank run.

Suppose that the bank has chosen a portfolio (x, y) and a deposit contract d at date 0. At date 1 the budget constraint requires

$$\lambda d \leq y$$

and we can assume, without loss of generality, that the intermediary always chooses (x, y) and d to satisfy this constraint. Otherwise, the intermediary will always have to default and the value of d becomes irrelevant. Consequently, the consumption of the late consumers, conditional on no run, will be given by

$$(1 - \lambda)c_{2s} = R_s(1 - y) + y - \lambda d.$$

This is consistent with no run if and only if $c_{2s} \geq d$ or

$$d \leq R_s(1 - y) + y.$$

This last inequality is called the *incentive constraint*. If this inequality is satisfied, there is an equilibrium in which late consumers wait until date 2 to withdraw. Since we only admit essential runs, the necessary and sufficient condition for a bank run is that the incentive constraint is violated, that is,

$$d > R_s(1 - y) + y.$$

Since $R_H > R_L$, this condition tells us that there can never be an essential run in state H unless there is also one in state L. There is no point choosing d so large that a run always occurs, so we can restrict attention to cases in which a run occurs in state L if it occurs at all. There are then three different cases that need to be considered. In the first, the incentive constraint is never binding and bankruptcy is not a possibility. In the second case, bankruptcy is a possibility but the bank finds it optimal to choose a deposit contract and portfolio so that the incentive constraint is (just) satisfied and there is no bankruptcy in equilibrium. In the third case, the costs of distorting the choice of deposit contract and portfolio are so great that the bank finds it optimal to allow bankruptcy in the low asset-return state.

Case I: The incentive constraint is not binding in equilibrium

In this case, we solve the intermediary's decision problem without the incentive constraint and then check whether the constraint is binding or not. The intermediary chooses two variables, the portfolio y and the deposit contract d to maximize the depositor's expected utility, assuming that there is no bank run. With probability λ the depositor is an early consumer and receives d regardless of the state. With probability $1 - \lambda$, the depositor is a late consumer and then his consumption depends on the return to the risky asset. The total consumption in state s is equal to the return to the risky asset plus the remainder of the returns from the safe asset after the early consumers have received their share, that is, $R_s(1 - y) + y - \lambda d$. The consumption of a typical late consumer is just this amount divided by the number of late consumers $1 - \lambda$. Thus, the expected utility is

$$\lambda U(d) + (1-\lambda)\left\{\pi_H U\left(\frac{R_H(1-y)+y-\lambda d}{1-\lambda}\right)\right.$$
$$\left.+\pi_L U\left(\frac{R_L(1-y)+y-\lambda d}{1-\lambda}\right)\right\}.$$

This expression is maximized subject to the feasibility constraints $0 \le y \le 1$ and $\lambda d \le y$.

Assuming that the optimal portfolio requires investment in both assets, i.e. $0 < y < 1$, the optimal choice of (y, d) is characterized by the necessary and sufficient first-order conditions. Differentiating the objective function with respect to d and taking account of the constraint $\lambda d \le y$, the first-order condition for the deposit contract is

$$U'(d) - \left\{\pi_H U'\left(\frac{R_H(1-y)+y-\lambda d}{1-\lambda}\right) + \pi_L U'\left(\frac{R_L(1-y)+y-\lambda d}{1-\lambda}\right)\right\} \ge 0,$$

with equality if $\lambda d < y$. Differentiating with respect to y and taking account of the constraint $\lambda d \le y$, the first-order condition for the portfolio is

$$\pi_H U'\left(\frac{R_H(1-y)+y-\lambda d}{1-\lambda}\right)(1-R_H) + \pi_L U'\left(\frac{R_L(1-y)+y-\lambda d}{1-\lambda}\right)(1-R_L) \le 0,$$

with equality if $\lambda d < y$. If (y^*, d^*) denotes the solution to these inequalities, then (y^*, d^*) represents an equilibrium if the incentive constraint is satisfied in state L:

$$d^* \le R_L(1-y^*)+y^*.$$

Let U^* denote the maximized value of expected utility corresponding to (y^*, d^*).

The consumption profile offered by the bank is illustrated in Figure 3.6, which shows the consumption at each date and in each state as a function of the return on the long asset. If the low state return R_L is sufficiently high, say $R_L = R_L^*$, then the incentive constraint is never binding. The early consumers receive $c_{1s} = d = y/\lambda$ and the late consumers receive $c_{2s} = R_s\left(1-y\right)/(1-\lambda)$ in each state $s = H, L$.

If the solution to the relaxed problem above does not satisfy the incentive constraint, there are two remaining possibilities: either the intermediary chooses a contract that satisfies the incentive constraint, i.e. one that is constrained by it, or the intermediary chooses a contract that violates the

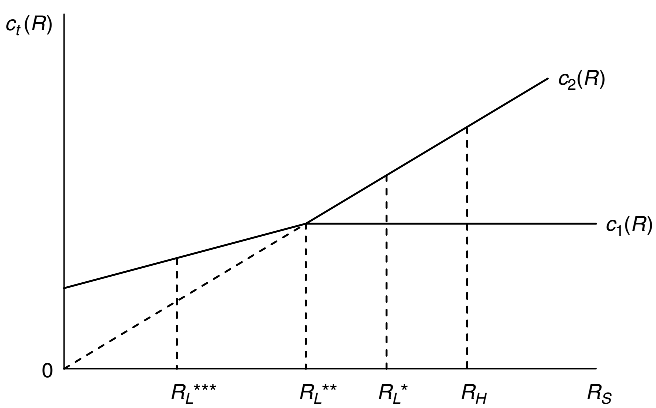

Figure 3.6. Illustration of consumption in each period and state as a function of the long asset's return R_s for $s = H, L$.

incentive constraint in the low state, in which case there is default in the low state.

Case II: The incentive constraint is binding in equilibrium

Suppose that (y^*, d^*) does not satisfy the incentive constraint. If the intermediary chooses not to default, the decision problem is to choose (y, d) to maximize

$$\lambda U(d) + (1-\lambda)\{\pi_H U(c_H) + \pi_L U(c_L)\}$$

subject to the feasibility constraints $0 \le y \le 1$ and $\lambda d \le y$ and subject to the incentive constraints

$$c_{2s} = \frac{R_s\left(1-y\right) + y - \lambda d}{1-\lambda} \ge d, \text{ for } s = H, L.$$

The incentive constraint will only bind in the low states $s = L$. Substituting for $c_{2L} = d$, the expression for expected utility can be written as

$$\lambda U(d) + (1-\lambda)\left\{\pi_H U\left(\frac{R_H(1-y) + y - \lambda d}{1-\lambda}\right) + \pi_L U(d)\right\},$$

where the assumption of a binding incentive constraint implies that

$$d \equiv R_L(1-y) + y.$$

Since d is determined by the choice of y, the optimal contract is entirely determined by a single first-order condition. Substituting for d in the objective function we obtain

$$\{\lambda + (1-\lambda)\pi_L\}U(R_L(1-y)+y) \\ + (1-\lambda)\pi_H U\left(\frac{(R_H - \lambda R_L)(1-y) + (1-\lambda)y}{1-\lambda}\right)$$

Note that the feasibility condition $\lambda d \le y$ must also be satisfied. We treat this as a constraint while maximizing the objective function above. Then the first-order condition for y takes the form

$$\{\lambda + (1-\lambda)\pi_L\}U'(d)(1-R_L) \\ + (1-\lambda)\pi_H U'\left(\frac{(R_H - \lambda R_L)(1-y) + (1-\lambda)y}{1-\lambda}\right)\left(\frac{-R_H + \lambda R_L + 1 - \lambda}{1-\lambda}\right) \le 0$$

or

$$(\lambda + (1-\lambda)\pi_L)U'(d)(1-R_L) \\ + \pi_H U'\left(\frac{R_H(1-y)+y-\lambda d}{1-\lambda}\right)(-R_H+\lambda R_L+1-\lambda) \leq 0,$$

with equality if $\lambda d < y$.

Let (y^{**}, d^{**}) denote the solution to this problem and let U^{**} denote the corresponding maximized expected utility.

This case is also illustrated in Figure 3.6. If $R_L = R_L^{**}$ then the incentive constraint is binding and consumption is the same for early and late consumers in the low state: $c_{1L} = c_{2L} = d = y + R_L\left(1-y\right)$. In the high state, $c_{1H} = d$ and $c_{2H} = R_H\left(1-y\right)/\left(1-\lambda\right)$ as usual.

Case III: The incentive constraint is violated in equilibrium

Again, suppose that (y^*, d^*) does not satisfy the incentive constraint. If there is default in the low state, the expected utility of the depositors is

$$\pi_H\left\{\lambda U(d) + (1-\lambda)U\left(\frac{R_H(1-y)+y-\lambda d}{1-\lambda}\right)\right\} + \pi_L U\left(r(1-y)+y\right).$$

In this case, the first-order conditions that characterize the choice of d and y take the form

$$\pi_H\left\{\lambda U'(d) - \lambda U'\left(\frac{R_H(1-y)+y-\lambda d}{1-\lambda}\right)\right\} \geq 0,$$

with equality if $\lambda d < y$ and

$$\pi_H U'\left(\frac{R_H(1-y)+y-\lambda d}{1-\lambda}\right)(1-R_H) + \pi_L U'\left(r(1-y)+y\right)(1-R_L) \leq 0,$$

with equality if $\lambda d < y$. Let $\left(d^{***}, y^{***}\right)$ denote the solution to this problem and U^{***} the corresponding maximized expected utility.

This case is illustrated in Figure 3.6. If $R_L = R_L^{***}$ then bankruptcy occurs in the low state and both early and late consumers receive the same consumption: $c_{1L} = c_{2L} = y + R_L\left(1-y\right) < d$. In the high state, $c_{1H} = d$ and $c_{2H} = R_H\left(1-y\right)/\left(1-\lambda\right)$ as usual. This is an equilibrium solution only if

$$d^{***} > R_L(1-y)+y,$$

and

$$U^{***} > U^{**}.$$

The first condition guarantees that the incentive constraint is violated, so that the intermediary must default in state L, and the second condition guarantees that default is preferred to solvency. Otherwise the bank prefers (d^{**}, y^{**}) and there is no default.

3.8 THE GLOBAL GAMES APPROACH TO FINDING A UNIQUE EQUILIBRIUM

Section 3.6 demonstrated how the sunspot approach allowed a complete description of an equilibrium with bank runs. The weakness of this approach is that it does not explain why the sunspot should be used as a coordination device. There is no real account of what triggers a crisis. This is particularly a problem if there is a desire to use the theory for policy analysis.

Carlsson and van Damme (1993) showed how the introduction of a small amount of asymmetric information could eliminate the multiplicity of equilibria in coordination games. They called the games with asymmetric information about fundamentals *global games*. Their work showed that the existence of multiple equilibria depends on the players having common knowledge about the fundamentals of the game. Introducing noise ensures that the fundamentals are no long common knowledge and thus prevents the coordination that is essential to multiplicity. Morris and Shin (1998) applied this approach to models of currency crises. Rochet and Vives (2004) and Goldstein and Pauzner (2005) have applied the same technique to banking crises. In this section we present a simple example of the global games approach provided by Allen and Morris (2001).

There are two depositors in a bank. Depositor i's type is ℓ_i. If ℓ_i is less than 1, then depositor i is an early consumer and needs to withdraw his funds from the bank. If ℓ_i is greater than or equal to 1, he is a late consumer and has no liquidity needs. In this case he acts to maximize his expected return. If a depositor withdraws his money from the bank, he obtains a guaranteed payoff of $\omega > 0$. If both depositors keep their money in the bank then both obtain ρ where

$$\omega < \rho < 2\omega.$$

If a depositor keeps his money in the bank and the other depositor withdraws, he gets a payoff of 0.

Note that there are four states of liquidity demand: both are early consumers and have liquidity needs, depositor 1 only is an early consumer and has liquidity needs, depositor 2 only is an early consumer and has liquidity needs, and both are late consumers and have no liquidity needs. If there is common knowledge of fundamentals, and at least one depositor is an early consumer, the unique equilibrium has both depositors withdrawing. But if it is common knowledge that both depositors are late consumers, they are playing a coordination game with the following payoffs. (The first element represents the payoff to the player choosing the row strategy and the second element is the payoff to the player choosing the column strategy.)

	Remain	Withdraw
Remain	(ρ, ρ)	$(0, \omega)$
Withdraw	$(\omega, 0)$	(ω, ω)

An important feature of this coordination game is that the total payoffs when only one person withdraws early are less than when both people withdraw early. One set of circumstances where this would arise, for example, is when the bank can close down after everybody has withdrawn, but when anybody keeps their money in the bank then extra costs are incurred to stay open and the bank's assets are dissipated more.

With common knowledge that neither investor is an early consumer, this game has two equilibria: both remain and both withdraw. We will next consider a scenario where neither depositor is an early consumer, both know that no one is an early consumer, both know that both know this, and so on up to any large number of levels, but nonetheless it is not common knowledge that no one is an early consumer. We will show that in this scenario, the unique equilibrium has both depositors withdrawing. In other words, beliefs about others' beliefs, or higher-order beliefs as they are called, in addition to fundamentals, determine the outcome.

Here is the scenario. The depositors' types, ℓ_1 and ℓ_2, are highly correlated; in particular suppose that a random variable T is drawn from a smooth distribution on the non-negative numbers and each ℓ_i is distributed uniformly on the interval $[T - \varepsilon, T + \varepsilon]$, for some small $\varepsilon > 0$. Given this probability distribution over types, types differ not only in fundamentals, but also in beliefs about the other depositor's fundamentals, and so on.

To see why, recall that a depositor is an early consumer if ℓ_i is less than 1. But when do both depositors know that both ℓ_i are greater than or equal to 1 so they are late consumers? Only if both ℓ_i are greater than $1 + 2\varepsilon$. This is because both players knows that the other's ℓ_i is within 2ε of their own. For example, suppose $\varepsilon = 0.1$ and depositor 1 has $\ell_1 = 1.1$. She can deduce that T is within the range 1.0–1.2 and hence that ℓ_2 is within the range 0.9–1.3. Only if $\ell_1 \geq 1.2$ does depositor 1 know that depositor 2 is a late consumer.

When do both investors know that both investors know that both ℓ_i are greater than or equal to 1? Only if both ℓ_i are greater than $1 + 4\varepsilon$. To see this, suppose that $\varepsilon = 0.1$ and depositor 1 receives $\ell_1 = 1.3$. She can deduce that T is within the range 1.2–1.4 and hence that depositor 2's signal is within the range 1.1–1.5. However, if depositor 2 receives $\ell_2 = 1.1$, then he sets a positive probability of depositor 1 having ℓ_1 within the range 0.9–1.3 as above. Only if depositor 1's signal is greater or equal to $1 + 4\varepsilon$ would this possibility be avoided and both would know that both know that both are late consumers.

As we go up an order of beliefs the range goes on increasing. Hence it can never be common knowledge that both depositors are late consumers and have no liquidity needs.

What do these higher-order beliefs imply? It is simplest to consider what happens when ε is very small. In this case, since T is smoothly distributed the probability of the other depositor having an ℓ_i above or below approaches 0.5 in each case as $\varepsilon \to 0$ (see Figure 3.7). We will take it as 0.5 in what follows. (An alternative approach is to assume T is uniformly distributed in which case it is exactly 0.5 even away from the limit of $\varepsilon \to 0$ – see Morris and Shin (2003).)

How do depositors behave in equilibrium? Observe first that each depositor will withdraw if $\ell_i < 1$ so the depositor is an early consumer. What about if $\ell_i \geq 1$? Given the structure of the model with a person being an early consumer when $\ell_i < 1$ and a late consumer when $\ell_i \geq 1$, the most natural strategy for a depositor to follow is to choose a strategy of remaining only when $\ell_i > k$

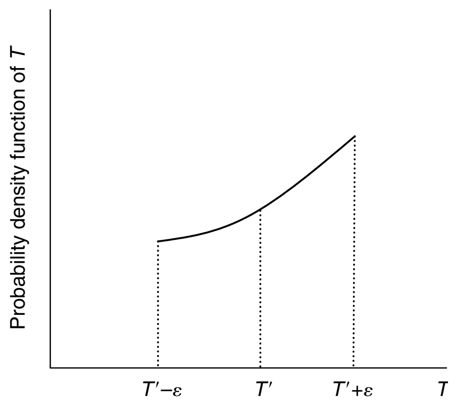

Figure 3.7. The probability of the other depositor's ℓ_i being above or below 0.5 as $\varepsilon \longrightarrow 0$.

for some $k \geq 1$ and withdrawing otherwise. Suppose depositor 1 follows this strategy. Consider what happens when $\ell_2 = k$. Given our assumptions about ε being small and T being drawn from a smooth distribution, depositor 2 deduces that there is a 0.5 probability that $\ell_1 < k$ and depositor 1 will withdraw and a 0.5 probability that $\ell_1 \geq 1$ and that she will remain. The payoff of depositor 2 from **remaining** is

$$= 0.5 \times \rho + 0.5 \times 0 = 0.5\rho,$$

and the payoff from **withdrawing** is

$$= 0.5 \times \omega + 0.5 \times \omega = \omega.$$

Since it is assumed that $\rho < 2\omega$ or equivalently that $0.5\rho < \omega$ it follows that depositor 2 will also withdraw. In fact his unique best response is to withdraw if ℓ_2 is less than some cutoff point k^* strictly larger than k where at k^* the expected payoffs from remaining and withdrawing are equated. Since the two depositors are in symmetric positions, we can use the same argument to show that depositor 1 will have a cutoff point higher than k^*. There is a contradiction and both remaining cannot be an equilibrium. In fact the equilibrium for small ε is unique, with both agents always withdrawing. Given the other is withdrawing, it is always optimal to withdraw.

The argument ruling out the equilibrium with both remaining depends on the assumption $\rho < 2\omega$. If this inequality is reversed then the same logic as above can be used to show that the unique equilibrium has both remaining. Again the multiplicity is eliminated.

The arguments used above to eliminate an equilibrium rely on depositors using a switching strategy where below some level they withdraw and above some level they remain. Note that other types of equilibria have not been ruled out here. For a full analysis of global games, see Morris and Shin (2003).

Using a global games approach to ensure the uniqueness of equilibrium is theoretically appealing. It specifies precisely the parameter values for which a crisis occurs and allows a comparative static analysis of the factors that influence this set. This is the essential analytical tool for policy analysis. However, what is really needed in addition to logical consistency is empirical evidence that such an approach is valid. Currently there is a very limited empirical literature. This is in the context of currency crises and is broadly consistent with the global games approach (see Prati and Sbracia 2002; Tillman 2004; and Bannier 2005).

3.9 LITERATURE REVIEW

There is a large literature on banking crises. Excellent surveys are provided by Bhattacharya and Thakor (1993), Freixas and Rochet (1997), and Gorton and Winton (2003). This review will therefore be brief.

Bryant (1980) and Diamond and Dybvig (1983) developed the first models of banking crises. Both papers had consumers with random liquidity demands and showed that deposit contracts allowed this risk to be insured. In Diamond and Dybvig (1983) bank runs were generated by sunspots while in Bryant (1980) they were the result of aggregate loan risk and asymmetric information about loan payoffs. Both papers were concerned with justifying deposit insurance. Diamond and Dybvig argue deposit insurance eliminates the bad equilibrium without cost because it removes the incentives of late consumers to withdraw early so in equilibrium there are no costs to providing the insurance. In Bryant's model, deposit insurance is desirable because it eliminates incentives to gather costly information that is not socially useful.

Following Diamond and Dybvig, much of the literature on panic-based runs was focused on the assumptions underlying their model. Cone (1983) and Jacklin (1987) pointed out that it was necessary for depositors to have restricted trading opportunities, otherwise banks would have to compete with financial markets and this would eliminate the insurance they could offer.

As we have seen, the possibility of panic-based bank runs depends in an important way on the sequential service (or "first come-first served") constraint. Without this, runs could be prevented by suspending convertibility. A number of papers sought to justify the existence of the sequential service constraint endogenously rather than appealing to legal restrictions. Wallace (1988) assumes that the fraction of the population requiring liquidity is random. He also assumes that agents are spatially separated from each other but are always in contact with the bank. These factors imply that a sequential service constraint is optimal. Whereas Diamond and Dybvig were able to show that deposit insurance was optimal, in Wallace's model it is not. Building on Wallace's model, Chari (1989) shows that if the interbank market does not work well, because of regulatory restrictions of the type in place during the National Banking Era in the US, then banking panics can occur. With a well functioning interbank market, however, runs do not occur. Calomiris and Kahn (1991) argue that the deposit contract, together with a sequential service constraint, can be optimal because it provides an incentive for depositors to monitor bank managers. Diamond and Rajan (2001) show that the possibility of runs arising from demand deposits and a sequential service constraint can be desirable if it ensures that banks will not renegotiate to extract more rents from entrepreneurs that have borrowed from the bank.

The second category of crises involves aggregate risk arising from the business cycle. Bryant's (1980) model falls in this category since he assumes aggregate loan risk and asymmetric information about the outcome of this risk to produce an incentive for some depositors to run. In Gorton's (1985) model, depositors receive a noisy signal about the value of bank assets and if this suggests the value is low there is a panic. Banks that are solvent suspend convertibility and pay a verification cost to demonstrate this to investors. 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 unfavorable 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. As discussed above, Allen and Gale (1998) develop a model of business cycle risk with symmetric information where future prospects can be observed by everybody but are not contractible. Runs occur when future prospects are poor.

As Section 3.6 illustrated, one of the key issues in the bank-run literature is that of equilibrium selection. Diamond and Dybvig appealed (among other things) to sunspots to act as the coordination device but did not model this fully. Postlewaite and Vives (1987) developed a model that does not rely on sunspots and that generates a unique equilibrium. In the context of currency crises, Morris and Shin (1998) showed how the global games approach of Carlsson and van Damme (1993) could be used to ensure equilibrium is unique. Their approach links the panic-based and fundamental-based approaches by showing how the probability of a crisis depends on the fundamentals. Morris and Shin (2003) provide an excellent overview of global games. Allen and Morris (2001) develop a simple example to show how these ideas can be applied to banking crises. Rochet and Vives (2004) use the unique equilibrium resulting from their global games approach to undertake policy analysis. They consider the role of ex ante regulation of solvency and liquidity ratios and ex post provision of liquidity by the central bank. Goldstein and

Pauzner (2005) use the global games approach to show how the probability of panic-based runs can be made endogenous and related to the parameters of the banking contract.

There is a large empirical literature on banking crises, which will only be briefly touched on here. Sprague (1910) is the classic study of crises in the National Banking Era. It was commissioned by the National Monetary Commission after the severe crisis of 1907 as part of an investigation of the desirability of establishing a central bank in the US. Friedman and Schwartz (1963) have written a comprehensive monetary history of the US from 1867 to 1960. Among other things, they argue that banking panics can have severe effects on the real economy. In the banking panics of the early 1930's, banking distress developed quickly and had a large effect on output. Friedman and Schwartz argued that the crises were panic-based and offered as evidence the absence of downturns in the relevant macroeconomic time series prior to the crises. Gorton (1988) showed that banking crises in the National Banking Era were predicted by a leading indicator based on liabilities of failed businesses. This evidence suggests banking crises are fundamental or business cycle related rather than panic-based. Calomiris and Gorton (1991) provide a wider range of evidence that crises are fundamental-based rather than panic-based. Wicker (1980, 1996) shows that, despite the absence of collapses in US national macroeconomic time series, in the first two of the four crises identified by Friedman and Schwartz in the early 1930's there were large regional shocks and attributes the crises to these shocks. Calomiris and Mason (2003) undertake a detailed econometric study of the four crises using a broad range of data and conclude that the first three crises were fundamental-based while the fourth was panic-based.

3.10 CONCLUDING REMARKS

Banking crises have been an important phenomenon in many countries in many historical periods. In this chapter we have developed a framework based on Bryant (1980) and Diamond and Dybvig (1983) for analyzing these crises. There are two approaches that can be captured by the framework. The first is crises that are based on panics. The second is crises that are based on poor fundamentals arising from the business cycle. There has been a significant debate in the literature on which of these is the "correct" approach to take to crises. As we have seen, there is evidence that both are empirically relevant. There is no need to confine attention to one or the other as is done in much of the literature. Both are important.