The Scandinavian
Journal of Economics

Scand. J. of Economics 125(3), 553–583, 2023
DOI: 10.1111/sjoe.12535

Banks and financial crises: contributions of Ben Bernanke, Douglas Diamond, and Philip Dybvig*

Zhiguo He

University of Chicago, Chicago, IL 60637, USA
zhiguo.he@chicagobooth.edu

Yunzhi Hu

University of North Carolina, Chapel Hill, NC 27599, USA
yunzhi_hu@kenan-flagler.unc.edu

Abstract

The 2022 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel was awarded to Ben S. Bernanke, Douglas W. Diamond, and Philip H. Dybvig "for research on banks and financial crises". This article surveys the contributions of the three laureates and discusses how their insights have changed the way that academics and policymakers understand banks and their roles in financial crises.

Keywords: Bank monitoring; bank regulation; bank runs; financial accelerator; financial crises; macro finance; monitoring

JEL classification: *E*51; *E*53; *G*21; *G*28

1. Introduction

The 2022 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel was awarded to Ben S. Bernanke, Douglas W. Diamond, and Philip H. Dybvig "for research on banks and financial crises". The prize committee mentioned three papers by the laureates, all published in the early 1980s. However, both banks and financial crises have a much longer history, and it was not until the early 1980s that academics gave a satisfying answer to fundamental questions such as why banks exist, why they take specific forms in practice, and why bank failures can have detrimental effects on the macroeconomy. The work of the three laureates was primarily motivated by the Great Depression in the 1930s. Their contributions, in turn, have significantly shaped the policies of central

*We are grateful for comments from Yasser Boualam, Douglas Diamond, Raghuram Rajan, Pontus Rendahl, Roberto Robatto, and Marianne Simonsen.

banks and governments around the globe during the 2007–2008 global financial crisis.

The world's oldest bank is Banca Monte dei Paschi di Siena (MPS). MPS was initially established as a mount of piety and took its present form in 1624. The first documented bank run occurred in 1866, when Overend, Gurney and Company, a London wholesale discount bank, suspended payments and had large crowds around the head offices (Sowerbutts et al., 2016). Before 2007–2008, the United States experienced patterns that might arguably be classified as financial crises in 1797, 1814, 1819, 1825, 1833, 1837, 1857, 1861, 1864, 1873, 1884, 1890, 1893, 1907, and 1914 (Gorton and Tallman, 2018). Hammond (1991) provides a historical view of how banking evolved in the United States in the context of the nation's political and social development. The 1929–1939 Great Depression put banks at the focus of the macroeconomy discussion. Triggered by the 1929 stock market crash, depositors began to panic and looked for safe storage of their physical cash. The first bank run kicked off in 1930 in Nashville, Tennessee,[1] which eventually led to the failure of 9,000 banks and wiped off $7 billion in depositors' wealth.[2] If banks are so frequently associated with financial crises that are detrimental to the macroeconomy, why do they exist? Modern banks take a two-layer structure: depositors, such as households, invest their money in banks, which in turn lend to borrowers such as entrepreneurs. Why is such a two-layer structure necessary? Why don't households directly lend to entrepreneurs? Moreover, why does the contract between households and the bank take the format of a deposit contract, which allows the households to withdraw their money as they wish? In contrast, why does the contract between the bank and entrepreneurs takes the form of a more standard debt contract specifying the payment amount and date? Finally, why are banks vulnerable to the risk of bank runs by depositors, why are runs so harmful to the macroeconomy, and how can bank regulation and other government policies, such as monetary policy, mitigate runs?

In this paper, we survey the academic contribution made by the three laureates. All three laureates are prolific and, given the constraints, we will have to restrict ourselves to their academic papers that directly relate to banks and financial crises. These papers include both theoretical and empirical works. We will have to omit the other lines of work by the laureates.[3] Moreover, this paper does not aim to survey the enormous body of literature

[1] See https://www.history.com/topics/great-depression/bank-run.

[2] See https://www.ssa.gov/history/bank.html.

[3] Specifically, we omit Bernanke's work on investment cyclicality and monetary policy, Diamond's research on information efficiency, and Dybvig's study on portfolio choice and capital structure.

on banks and financial crises. For that purpose, we refer readers to the handbook chapter by Gorton and Winton (2003) and the textbook by Freixas and Rochet (2008). Instead, we focus on the contribution of the three laureates and we try to organize their work around the questions listed above. We also offer an introduction to understanding banks and financial crises before the ground-breaking work of the three laureates, and some later follow-up research.

The development of banking theory took off in the early 1980s. Even at its onset, researchers noted that banks' asset and liability sides are closely interconnected. It can be misleading to draw conclusions from exclusively studying one side of the bank's balance sheet. That said, the theoretical models still have a relative focus on either the asset or the liability side. In Section 2, we review the theories on the asset side of banks' balance sheets. We explain why bank assets are mostly loans, essentially debt contracts. Moreover, we show that banks' ability to monitor and diversify risks by lending to a large number of borrowers makes them unique, and leads to the two-layer banking structure widely observed in practice. A related question is, why do some firms prefer to borrow from banks, whereas others prefer to issue bonds and commercial papers in the public market? Why does bank debt tend to be short term and often more senior in a firm's capital structure than market debt?

In Section 3, we turn to theories that focus relatively more on the liability side of banks. Bank liabilities are mostly deposits, which allow depositors to withdraw as they wish under a first-come, first-served sequence. Deposits differ from standard debt contracts, with the latter specifying a principal amount and due dates. They are mostly issued by banks and other bank-like institutions. Why do banks issue deposits? What is the fundamental economic problem that deposits solve? Moreover, what is the role of equity in banks' capital structure?

In Section 4, we describe how standard macroeconomic models started to incorporate credit market frictions and banks in the late 1980s. Specifically, we introduce the well-known financial-accelerator theory, which shows how financial frictions in the credit market can amplify and prolong shocks to the macroeconomy, which could result in financial crises. Moreover, how do monetary policy shocks transmit to the real economy through these credit market frictions? Finally, given that financial crises are typically accompanied by a liquidity shortage – that is, liquidity supply falls below the demand – how do banks affect the liquidity supply that may directly contribute to crises?

In Section 5, we briefly discuss the policy implications from the laureates' research on how the government should assist banks during crises and regulate banks during normal times. We talk about policies such as the Federal Deposit Insurance Corporation (FDIC) insurance, the Federal Reserve's role as lender of last resort, and the liquidity regulations introduced by the Basel Committee after the 2007–2008 global financial crisis. In Section 6, we very briefly summarize some follow-up work that further extends the literature in different directions. Finally, just as we were due to submit the final version of our paper, Silicon Valley Bank (SVB) experienced a classical bank run and subsequently failed on 10 March 2023. SVB is the largest bank that has failed since the global financial crisis, and we provide a very brief account in Section 7.

2. The asset side of banks

Loans are the most common asset on a bank's balance sheet. They can be commercial and industrial loans, mortgages, car loans, student loans, credit cards, etc. In practice, the details of loan contracts can become fairly complicated: a standard credit agreement can easily extend to hundreds of pages and may involve thousands of terms. However, loans are essentially debt contracts that specify a fixed schedule of payments that borrowers need to repay to receive some upfront borrowing for investments into fixed assets or working capital. Penalties will be triggered whenever the borrower fails to make the payment. Why are loans, or general debts, so prevalent in the real world?

A natural answer is that among any possible contract between a borrower and her lenders, debt is the optimal one. Townsend (1979), who developed what is known as the costly state verification (CSV) model, was the first to show this result. In this paper, an entrepreneur has access to an investment opportunity but does not have enough wealth to invest, and thus needs to borrow. The investment is risky because it may sometimes generate very low cash flows. A crucial assumption is that the cash flows from investment cannot be directly assigned to investors because only the borrower observes them. In this situation, the borrower can always claim the realized cash flows are very low and divert away part of the actual cash flows for private consumption. Investors can pay a cost to audit; in this case, they will discover the true cash flows. The fundamental question is, therefore, how to offer incentives to the borrower to truthfully report the realized cash flows and repay investors without always triggering the audit cost. To the extent that auditing is costly, an optimal contract should try to minimize its occurrence. Townsend (1979) shows that the optimal contract is a debt contract. If the entrepreneur's reported cash flows exceed a threshold, investors do not audit but simply receive a constant payment, interpreted as the face value of a debt contract. If the entrepreneur's reported cash flows fall below the threshold, investors choose to audit and can punish the entrepreneur if she has lied.[4] The CSV model has become

[4] On the equilibrium path, the entrepreneur always truthfully reports, but investors must still commit to auditing if the reported cash flows are below the threshold.

the standard workhorse model for analyzing financial frictions in banking and macrofinance. This model explains why the contracts between borrowers and lenders often take the form of a debt contract. Moreover, it implies why borrowers need to have some personal “net worth” to get investment projects financed (see the discussion of Bernanke and Gertler, 1989, in Section 4.2). Specifically, as a result of the information asymmetry between borrowers and lenders, the optimal contract must entail some agency costs as deadweight losses relative to the first-best allocation without information asymmetry. These agency costs lead to more costly funding when the borrower seeks to borrow “externally”. When the borrower has a high net worth, the agency costs are lower, and she can use more of her own net worth (“internal” funds) to make investments.

Independently, the first part of Diamond (1984) studies a problem very similar to Townsend (1979). In this case, the lender can incur a costly non-pecuniary penalty (an interpretation is a legal cost in bankruptcy court). Here, the optimal contract is also a debt contract. Instead of auditing, Diamond (1984) introduces the option of monitoring. Specifically, a lender can spend some resources to observe the realization of the cash flows. Monitoring differs from auditing (Townsend, 1979) in that the cost of monitoring is incurred *ex ante*, before the cash flows are realized. In contrast, the cost of auditing is incurred *ex post*, after the cash flows are realized, and therefore it is state-contingent.[5]

The CSV model studies the problem of direct financing (i.e., one investor directly lending to one borrower). In practice, banks are often involved with indirect (or intermediated) financing. In particular, the business model of modern banks works as a two-layer structure. Investors deposit their money into a bank, which in turn lends to firms to invest in different projects. Why do banks take such a two-layer structure? What is the fundamental problem that banks are designed to solve? The second part of Diamond (1984) answers these questions. It starts with the observation that project investment typically incurs a large and fixed amount of money, and each individual investor’s wealth is insufficient to finance the fixed amount of borrowing. Therefore, a firm needs to borrow from a large set of investors. Each individual investor can monitor the firm, but monitoring efforts can be repetitive, which is inefficient. Meanwhile, monitoring has the property of public goods so that investors can free-ride each other. This can lead to an outcome where nobody monitors. To avoid the duplication of effort and the free-rider problem, naturally, monitoring should be delegated to one agent, who becomes the banker. An immediate follow-up question is, who monitors the monitor? Indeed, given

[5] Therefore, the monitoring decision does not require the lender’s commitment, whereas auditing does.

that only the bank and the borrowing firm observe the realized cash flows, the two parties have incentives to collude and cheat the remaining investors. The second part of Diamond (1984) provides an answer to the question of who monitors the monitor. The paper shows that when the bank is diversified by holding a large number of loans, the realized value of its portfolio is fairly predictable. In this case, investors in the bank can hold deposits – also a form of debt contract – as the optimal financing contract. If all the risks can be perfectly diversified, then a bank that monitors all its loans can finance all its lending using riskless deposits. If there are still some residual risks that cannot be diversified, the bank can use a combination of riskless deposits and risky claims such as equity. In both cases, there is no need to "monitor the monitor".

Diversified banks reflect the key idea of "financial engineering" in that they transform loans that need monitoring (informational sensitive) into deposits that do not (informational insensitive). Banks in Diamond (1984) shall be interpreted more broadly, including the securitization vehicles that conduct pooling (diversification) and tranching (selling off only senior claims). It is widely believed that these securitization vehicles played a central role in the 2007–2008 global financial crisis. The structure of pooling and tranching was later formalized by DeMarzo (2005), who studied the problem of an informed originator trying to sell assets to uninformed investors. DeMarzo shows that pooling multiple assets together to sell (such as selling a bundle) is dominated by selling assets individually. However, if there is sufficient diversification in the pool, then pooling (multiple assets together) with tranching can be the optimal arrangement.

2.1. Banks and firms' cost of capital

A closely related question on the asset side of banking is, why do borrowers choose to borrow from banks? In Diamond (1984), no single investor has sufficient wealth to finance the investment. Diamond (1991b) offers another reason based on reputation. In an earlier work, Diamond (1989) established that borrowers who pay their debts over time acquire a better reputation (for example, a better credit rating), which becomes an asset they lose if they subsequently default. For borrowers without an established reputation, Diamond (1991b) shows that bank monitoring can substitute for it. Therefore, Diamond predicts a separation in which new borrowers without a long track record need to be monitored, while others who have always repaid such debt for a long enough time acquire a sufficiently good reputation to borrow directly without monitoring. For borrowers in the first group, their investment choices are monitored by banks, and the record of successful repayments helps future lenders learn about their underlying business quality.[6] For borrowers in the second group, they can issue debt directly to public markets. Diamond also produces a life-cycle theory of borrowing: young borrowers (small and medium-sized businesses) borrow from banks, and mature borrowers with a good enough credit rating switch to unmonitored borrowing and no longer depend on bank finance.

Besides the choice between bank and market debt, another relevant question for firms is determining their debt's maturity. In practice, bank loans have, on average, shorter maturities than publicly issued bonds. Diamond (1991a) makes an important contribution to understanding debt maturities. He shows that the optimal debt maturity trades off the signaling role of short-term debt and its liquidity risk. Specifically, Diamond assumes that the firm's management has private information about the firm's credit risks, whereas creditors only have access to some less precise indicators, such as credit ratings. For all firms with the same credit rating, some are more creditworthy than others. Given this fact, the more creditworthy borrowers will have incentives to signal themselves. In Diamond (1991a), short-term debt fulfills such a signaling role. The management's private information today will gradually become public over time. When lenders observe such good information, they update their beliefs about a firm's credit risk and, consequently, reduce the spreads charged. In other words, short-term debt enables firms' borrowing costs to be more sensitive to public information as the information becomes available to lenders. On average, firms with low risks are more likely to generate positive public information than those with high risks. Hence, using short-term debt is generally beneficial for low-risk firms. However, refinancing short-term debt creates liquidity risk. Even low-risk firms may generate negative news to the public, such as temporarily low profits. In this case, these firms could find it difficult or even impossible to refinance short-term debt – a type of liquidity risk. Diamond (1991a) predicts that firms with both high and low credit ratings prefer short-term debt, whereas firms with intermediate ratings prefer long-term debt. For high-rated firms, the liquidity risk is low. For low-rated firms, the lenders will only extend short-term debt.

Besides having shorter maturities, bank loans are typically senior to public bonds in firms' capital structure. That is, in the case of bankruptcy, the priority is to repay bank loans before public bonds. Diamond (1993) offers an explanation and shows why short-term debt should generally be senior to long-term debt. The insights are closely related to those in Diamond (1991a). In the model, borrowers have private information about the prospect of their future credit ratings. Short-term debt allows the borrower to refinance and

[6] Hu and Varas (2021) show that when the repayment record is not observable by future lenders, the bank has incentives to engage in zombie lending.

reduce the cost of credit after they obtain positive information. However, it can lead to excessive liquidation because lenders ignore borrowers' control rents. Making short-term debt senior to long-term debt increases the firm's overall financing cost sensitivity to new public information for a given liquidation level. Following bad news, long-term debt will allow the issuance of short-term debt even if this dilutes the value of long-term debt. This dilution can prevent the liquidation of solvent but illiquid firms, which is socially efficient. Diamond (1993) implies that financial intermediaries will hold short-term senior debt, whereas the public will hold long-term junior debt.

A closely related issue to financial crises is why firms, both financial and non-financial, borrow so much short-term debt. Although more detailed answers will be provided in the next section, where we discuss the liability side of banks, the general economic mechanism, as highlighted in Brunnermeier and Oehmke (2013), is related to the idea that short-term debt enjoys a higher effective seniority over long-term debt simply because the former is paid earlier than the latter.[7] However, because seniority of the debt contract is the driving force of debt overhang, the above logic would suggest that short-term debt should impose a greater overhang than long-term debt, contradicting one of the key takeaways from the classic Myers (1977). By analyzing several workhorse frameworks in corporate finance, Diamond and He (2014) point out that short-term debt could impose stronger debt overhang if the firm's underlying assets exhibit counter-cyclical (stochastic) volatility – for instance, loans with higher volatility in bad times than in good times. Because banks typically retain loans as assets on their balance sheet, this result is particularly interesting in the sense that, compared with non-financial firms, short-term debt imposes more debt overhang for financial firms.

Diamond and Verrecchia (1991) study how a firm's future disclosure policy reduces its cost of capital today. Specifically, they study large investors who have private information on the firm but who may also experience future liquidity shocks, in which case they must sell their shares. A commitment to disclosure reduces information asymmetries in the future, which encourages large investors to hold the firm's shares today. This increases the prices today and hence lowers the costs of issuing capital.

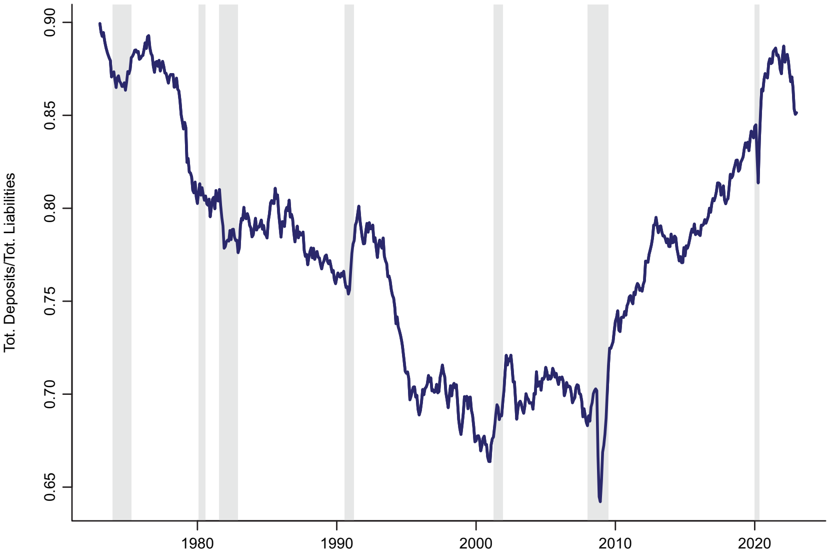
3. The liability side of banks

3.1. Demand deposits

As shown in Figure 1, deposits are the most common type of liability on the bank's balance sheet. Deposits are a special form of debt. Some of them are

[7] Hu et al. (2021) show that this result holds even without explicit seniority structure in bankruptcy.

Figure 1. Deposits to liabilities of banks in the United States: January 1975 to January 2023



Notes: This figure plots the time series of total deposits to total liabilities of all commercial banks in the United States.

demandable, meaning they can be withdrawn or accessed by depositors at any time.[8] This contrasts with standard debt contracts with a fixed maturity date and may incur penalties for early withdrawal. Why do banks issue deposits? What is the fundamental economic problem that demand deposits can solve?

These questions are studied in the influential paper by Diamond and Dybvig (1983). Many people, including academics, think the main contribution of Diamond and Dybvig (1983) is to explain bank runs. This is not entirely true. The phenomenon of bank runs and the mechanism behind them were pointed out earlier, even though they were not formalized.[9] Instead, the real contribution of Diamond and Dybvig (1983) is to show that banks can be the optimal solution to the problem of consumer uncertainties about their preferences for when to consume. In this model, consumers might need to

[8] An exception is time deposits, also known as certificates of deposit. Time deposits require the account holder to deposit for a fixed period of time and cannot be withdrawn without incurring a penalty.

[9] The famous movie *It's a Wonderful Life* in 1946 shows that panics can drive bank runs as opposed to anything fundamentally problematic.

consume early (such as in a medical emergency) or late, and they do not know the exact timing of their future consumption when making investment decisions. This is the notion of liquidity risk. They have some endowments, which can be either put in a short technology (such as storage) or a long technology that generates more in the long run but the same as the short technology in the near future. Naturally, risk-averse consumers would hope to have more ability to consume if they turned out early, even if this comes at the expense of lower long-term consumption if they turned out late. This is the notion of liquidity insurance and liquidity creation. If consumers self-provided liquidity, the outcome would be inefficient. In that case, if they turn out to prefer early consumption, they wish they could consume more than just the amount produced by the short technology and the liquidation from the long technology. If they turn out to prefer late consumption, they regret the investment decision and wish they had invested all endowments into the long technology. A financial market can allow the two types of consumers to trade after they learn their preferences and to improve their expected welfare. However, it still does not implement the socially optimal solution. The reason for this is that the financial market cannot offer insurance that depends on the consumers' consumption preference, which is private information. By contrast, banks can offer liquidity insurance by pooling these liquidity needs together. The number of consumers who need early consumption is almost deterministic among a large population. By pooling these liquidity needs together, the bank can promise to pay consumers who turn out early more than the technology will generate, and to pay those who turn out late less than the long technology will generate. Demand deposit contracts can be used to implement the first-best allocation and to provide the optimal amount of liquidity insurance. Unfortunately, demand deposits can expose banks to runs. If, for whatever reason, depositors lose confidence and panic, then everyone decides to withdraw early. In this case, the bank is committed to paying more than the early liquidation value of its assets and has nothing left for late consumers. Therefore, it does indeed make sense for everyone to withdraw early. As Diamond often puts it, the "fear of fear itself" or self-fulfilling prophecy can trigger runs. The Diamond–Dybvig model justifies a role for the government with taxation authority to offer deposit insurance or to provide discount window loans to remove the panic equilibrium. These policies are further discussed by Diamond and Dybvig (1986), who specifically argue that market discipline on banks, such as limiting deposit insurance or requiring banks to have uninsured subordinated short-term debt, can destabilize banks. Moreover, banks should not use insured deposits to fund entry into new lines of business that are significantly riskier, such as real estate speculation and equity underwriting. These activities contribute little to liquidity creation but oftentimes lead to excessive risk-taking. The proposal of 100 percent reserve banking would prevent banks from creating liquidity. Finally, Diamond and

Dybvig (1986) suggest that deposit insurance premiums should be based on the riskiness of the bank's loan portfolio.

The Diamond–Dybvig model has been very influential ever since the early 1980s. A particularly important question is whether the bank can offer contracts other than demand deposits that can implement the first-best allocations and, in the meantime, avoid runs. Jacklin and Bhattacharya (1988) offer such an implementation via mutual funds. Instead of issuing demand deposits, the bank can issue all equity and pay dividends. Using dividends, late consumers can buy shares from early consumers.[10] Moreover, Jacklin (1987) shows that the banking solution cannot coexist with a competitive market solution. Under the optimal contracts offered by the bank, if there is also an anonymous market to trade the bank's demand deposits, an individual will find it optimal to invest in the long technology. This deviation implies that if the role of banks is to offer liquidity insurance, they cannot coexist with competitive markets. Diamond (1997) addresses this issue, and, in particular, introduces a financial market with limited participation.[11] Diamond (1997) responds to the critique of Jacklin (1987) by showing that banks can coexist with the financial market if the fraction of consumers who can participate in the financial market is not too high. Moreover, Diamond shows that limited participation in the financial market causes overinvestment in short-term real assets. However, banks can improve the liquidity that the financial market provides because long-term assets can be sold before maturity at higher prices than would prevail without banks. The model predicts that as households participate better in the financial market, the banking sector should shrink by holding fewer long-term assets. At the economy-wide level, however, more long-term assets are produced.

Diamond and Dybvig (1983) assume loans are illiquid. A natural question is why the bank cannot sell its loans to others (such as other banks) if depositors rush to withdraw early. A related question is, if loans are issued to fundamentally solvent borrowers, why can't the bank borrow against these loans? Moreover, demand deposits are paid first-come, first-served within a period (the so-called "sequential-service constraint"). In contrast, debts of non-financial firms are protected by the automatic stay and clawback provision prohibiting payments in anticipation of default. In Diamond and Dybvig (1983), the panic equilibrium is driven by this sequential service constraint, which introduces fragility. Why does the bank issue first-come,

[10] This mutual fund implementation is suboptimal under more general consumer preferences because it only equalizes the marginal rate of substitution between consumption on the two dates. The optimal allocation requires the marginal utilities to be equalized across types. Section 12.2 of Tirole (2010) provides a detailed analysis.

[11] To be precise, all early consumers can sell their assets in the financial market, but only a fraction of late consumers can buy these assets.

first-served deposits? Diamond and Rajan (2001) answer these questions. They show that the fragility of demandable deposits is not a bug, but a feature. This fragility is crucial for bankers to have the appropriate incentives to collect payments. They argue that loans are illiquid because payment collection requires the specific human capital of bankers (Hart and Moore, 1994). In other words, the banker is the relationship lender who has the capability to collect more payments relative to outside investors. This superior collection skill provides the banker with incentives to renegotiate down the payments to outside investors once the lending has been made. Diamond and Rajan (2001) show that the first-come, first-served demandable deposit can prevent the banker from initiating this renegotiation. Intuitively, whenever the banker proposes to reduce the payments, it is a dominating strategy for each depositor to run and collect the payments before others arrive at the bank. In other words, the sequential service constraint creates a collective action problem among depositors, which means that they make a run on the bank whenever they think their claim is in danger. The run causes the bank to fail, resulting in the banker losing any rent from control. Anticipating this, the banker never proposes renegotiation to begin with. Therefore, Diamond and Rajan (2001) conclude that demandable deposits create liquidity: they allow bankers to credibly commit to collect loans and repay to outside investors who are less skillful. Importantly, they conclude that the disciplinary role of demand deposits works only for financial intermediaries, not for direct borrowing by firms, explaining why banks are special.

A similar idea is developed in Diamond (2004), who studies the optimal financing contracts when the cost of enforcement is high. In emerging markets, lenders do not go to bankruptcy court after a borrower defaults; this is known as lender passivity because the cost of enforcing the contracts is high. Even though enforcement would serve to punish the borrower, it would also hurt the lenders due to the high enforcement cost. To convince lenders to go to court, the optimal arrangements should have the cost of enforcement borne by others – a type of externality. If the firm borrows from one lender, the lender should take senior debt and transfer the costs to equity holders. Borrowing from many creditors via short-term debt is another solution. In this case, the enforcement cost is imposed across different lenders. Once there is a contract violation, each creditor will run to the court and try to recover the payments, leaving the enforcement costs to others.

3.2. Bank run models before Diamond–Dybvig

Although bank runs are a recurring topic for economists and policymakers (see, e.g., the classic Aliber et al., 2015), Diamond and Dybvig (1983) is among the first group of academic papers that study bank runs rigorously and systematically. Perhaps more importantly, this literature was heavily influenced by Friedman and Schwartz (1963). Chapter 7 of that book, titled "The Great Contraction, 1929–33", provides a vivid account of the unprecedented turmoil experienced by the US financial system, including both the stock market and banks, during the Great Depression. As astutely put by Flood and Garber (1981), who heavily cite evidence and charts directly from Friedman and Schwartz (1963), "of course, in constructing a model of a systematic banking collapse, any researcher, at least half-seriously, probably is attempting an explanation of the enormous collapse of the great depression". This is perhaps why Bernanke (1983) spends the entire first section of that paper elaborating the background on how the financial system collapsed during the Great Depression, highlighting the correlation of the financial crisis with macroeconomic activities.

Before Diamond and Dybvig (1983), the leading researchers in this literature on bank runs put considerable emphasis on a different notion of "liquidity", often modeled as stochastic timing of payments (Patinkin, 1965), and "money", which is fiat currency in the sense of Samuelson (1958). Largely, it is because the profession often took banking services in practice as given. Bagehot's lender-of-last-resort rule – which aims to protect the money stock as the very first goal (see, e.g., Humphrey, 1986) – had a great impact on the way that researchers thought about this issue at that time.

There were several papers written along this line. Based on the framework of Samuelson (1958) and Bryant and Wallace (1980), Bryant (1980) presents a model in which demand deposits and the associated deposit insurance program can be beneficial. Demand deposits provide liquidity service to the so-called "early-diers"; these agents who suffer idiosyncratic preference shocks are similar to the early-type consumers in Diamond and Dybvig (1983). Clearly, demand deposits are just short-term claims of economic agents that are held against banks. To meet the redemption of these early-diers, banks in Bryant (1980) with such short-term liabilities need to keep some fractional currency reserves, a central topic that is analyzed in that paper.

Nevertheless, as stated by Bryant (1980, see the beginning of section 4 in the paper), "[while] the uninsurable risk introduced in the previous paragraph generates demand liabilities, this is not sufficient to produce bank runs. To generate bank runs, we add risky intermediary assets and asymmetric information. What is crucial for the bank runs is that the coexistence of the uninsurable risk of early death and the asymmetric information on the risky assets give the intermediary a signal-extraction problem." Although this idea captures an interesting economic mechanism and later is formally analyzed by Jacklin and Bhattacharya (1988), this modeling complication of information asymmetry unnecessarily obscures the "coordination" nature of bank runs as shown by Diamond and Dybvig (1983).

Of course, the "coordination" nature of bank runs was also explored by other researchers during that time. Flood and Garber (1981) study the endogenous timing of a systematic banking collapse under a deflationary environment. Because banks honor demand deposits as nominal liabilities while holding real assets whose nominal value shrinks over time along a (deterministic) deflationary path, the banking system will run out of business at the point when the nominal rates earned on the consol bonds exceed the bank's operating cost.

Interestingly, although the main model by Flood and Garber (1981) investigates a setting where market fundamentals drive a banking collapse, they also discuss the possibility of a collapse generated by mass hysteria, where the collapse of the banking system – or, equivalently, the shrinkage of the money multiplier – is caused by a self-fulfilling belief that a collapse will occur. Contrary to the prevailing view on bank runs at that time, this collapse – which occurs in the model in a predictable way – does not have to be either sudden or unanticipated. Conceptually, it is closely related to the coordination failure highlighted in the Diamond and Dybvig (1983) model.

It is worth highlighting that the paper by Diamond and Dybvig (1983) distinguishes itself from these works by taking one step back and deriving economic implications from the primitives. As explained there, by taking an optimal-contract approach under idiosyncratic "liquidity" preference shocks, Diamond and Dybvig (1983) show that banking itself can be thought of as an optimal insurance mechanism, and demonstrate that its associated bank runs – a form of bad equilibrium when implementing the optimal insurance mechanism by banking – display a general economic force of strategic complementarity. The banking solution and its associated bank runs are beyond the money setting, and they apply to an economy with real goods.

3.3. Bank equity

Whereas demand deposits create liquidity, they can also lead to runs and fragility. When the bank's asset risks are fully diversified, as in Diamond and Rajan (2001), there is no uncertainty regarding the final cash flows that the bank will collect. In this case, runs never occur (i.e., they are off the equilibrium path), and the bank finances itself using an all-deposit capital structure. With aggregate uncertainties that are observable but not verifiable, however, the all-deposit capital structure can lead to runs when asset values fall after bad aggregate shocks. In this case, it might be optimal for the bank to partially finance itself with a softer claim that can be renegotiated in bad times, hence the role of bank capital. Diamond and Rajan (2000) show that outside bank capital can mitigate runs but reduces liquidity creation. In their model, bank failure is costly because bankers lose their special expertise in collecting loan payments. Outside bank capital acts as a cushion against bank deposits, but its presence increases the rents that the banker can extract. There is an additional effect coming from bank capital. A capital-constrained bank with risks of failure can credibly threaten to liquidate borrowers and extract more payments from them.[12] Therefore, optimal bank capital structure trades off liquidity creation, the cost of bank failure, and the ability to extract payments from borrowers. Diamond and Rajan (2000) attribute the decline in bank capital to the improvement in the underlying investment project so that outside investors can collect more payments even without banks.

Bernanke and Gertler (1987) study the macroeconomic role of banks under frictions, as highlighted in Jensen and Meckling (1976). Banks can perform two intermediary functions. First, they evaluate and determine which projects are worthy of investment. Second, they audit the projects they have invested in to determine their true *ex post* returns. Both functions involve a fixed set-up cost, which explains bank specialization. In equilibrium, banks must hold a buffer, interpreted as bank capital, to guarantee the returns on their liabilities. The model of banking described above is embedded in a stylized general equilibrium framework. Bernanke and Gertler argue that banks matter to real activity mainly because they provide the only available conduit between savers and projects that require intensive evaluation and auditing. Factors that affect the ability and cost of the banking system to provide intermediation will therefore have an impact on the allocation in the real economy. These factors include the adequacy of bank capital, bank investments' riskiness, and bank monitoring costs.

4. Banking and financial crises

4.1. The Great Depression

The Great Depression in 1929 is probably the first financial crisis that academic scholars have carefully studied. For decades, researchers have debated the root cause of the Great Depression. Early works blame either overinvestment and overbuilding during the ebullient 1920s or the problem of "under-consumption" – the inability of households to purchase enough goods and services to utilize the economy's productive capacity (Bernanke, 2004). In 1963, Milton Friedman and Anna J. Schwartz published the classic book, *A Monetary History of the United States, 1867–1960*, which transformed the debate. Specifically, they offered evidence about the role of monetary factors, and they argued that "the [economic] contraction is, in fact, a tragic testimonial to the importance of monetary forces" (Friedman and Schwartz, 1963, p. 300). They also argued that the Federal Reserve had, up to that time, largely ignored the problems in the banking sector, which experienced enormous runs and failures between December 1930 and March 1933. They emphasized the

[12] The effect of bank capital on payment extraction is non-monotonic.

effects of bank failures on the money supply, which had detrimental effects on the broader economy.

Relatedly, another widely cited reason behind the Great Depression is the gold standard (Eichengreen and Sachs, 1985), a system in which the value of each currency is expressed in terms of ounces of gold. Under the gold standard, each participating country defines its monetary unit regarding a certain amount of gold. To maintain the gold standard, central banks had to promise to exchange actual gold for their paper currencies at the legal rate. Such promises exposed the system to speculative attacks, which led to deflation. Bernanke and James (1991) shows deflation and other constraints of the central banks due to the gold standard caused banking panics in both the United States and several other countries in the early 1930s. They argued that there might also be feedback whereby banking panics further intensified deflation, at least in the United States.

Bernanke (1983) argues that financial-market imperfections can play an important role in propagating and amplifying business-cycle fluctuations. In particular, Bernanke argues that the severity of the Great Depression was partly attributable to the loss in intermediary services suffered when the banking system collapsed in 1930–1933. In a nutshell, bank failures destroyed valuable banking relationships, reducing credit supply and harming the real economy. In contrast with Friedman and Schwartz (1963), Bernanke (1983) suggests that the Great Depression not only led to a contraction in the money supply but also had non-monetary effects on credit intermediation. These findings both motivate and confirm the role of banks and financial intermediaries in the macroeconomy.

So why are bank failures particularly detrimental to the real economy from the perspective of Bernanke (1983)? This is because the banking system provides critical credit intermediation to the real economy. Bernanke (1983, see section II) provides a "verbal model" – which serves as a useful theoretical framework – on this important conceptual point. In this model, the real service performed by the banking system is the ability to deal with information asymmetry (i.e., the differentiation between good and bad borrowers). More specifically, Bernanke (1983) defines the cost of credit intermediation to include the screening and monitoring costs of banks, who "develop expertise at evaluating potential borrowers, establishing long-term relationships with customers, and offering loan conditions that encourage potential borrowers to self-select in a favorable way". Then, it is easy to see that the problems in banking during 1930–1933 disrupted the credit allocation process. The cost of credit intermediation soared following the rapid switch away from banks. Of course, Bernanke (1983) also acknowledges the role of "bank runs" in the manner of Diamond and Dybvig (1983): "[the] fear of runs led to large withdrawals of deposits, precautionary increases in reserve-deposit ratios, and an increased desire by banks for very liquid or rediscountable assets. These factors, plus the actual failures, forced a contraction of the banking system's role in the intermediation of credit."

4.2. Financial accelerator theory

Most of the macroeconomic studies build on the conclusion from the classic paper Modigliani and Miller (1958) without carefully examining the underlying assumptions. The Modigliani–Miller theorem asserts that economic decisions do not depend on the financial structure in a setting with perfect capital markets. Under this perfect-capital market assumption, adding financial intermediaries to traditional macroeconomic models should have no consequence for real activity. However, the 2007–2008 global financial crisis highlighted the significant role that banks and other financial institutions play in the macroeconomy and how their actions can have wide- and far-reaching effects.

Bernanke and Gertler (1989) were among the first to introduce financing frictions into standard macroeconomic models. Their model builds on the CSV model in Townsend (1979) but allows the entrepreneurs to accumulate wealth. They show that the entrepreneurs' net worth matters for the real economy and temporary shocks to firms' internal resources can have a persistent effect on the aggregate output. An initial positive shock to the economy improves firms' profits and retained earnings; this, in turn, leads to increased investment and output, which amplifies the upturn. By contrast, a negative shock, albeit temporary, reduces the firms' profits and leads to decreased investment and output. Even though Bernanke and Gertler (1989) did not explicitly study banks, one can interpret the entrepreneurs in this model as bankers. Bernanke and Gertler (1990) build on the following observation: auditing costs – the agency costs in the CSV model – are empirically too small to rationalize first-order effects for financial fragility. Therefore, they build a more realistic model with asymmetric information on borrower types, borrower actions, and project qualities. The agency problem is that borrowers are insufficiently selective and may undertake negative present-value projects – identified by Jensen (1988) as the "free cash flow" problem. The model makes it easier to motivate quantitatively significant real effects for financial factors, as the empirical counterpart of the agency costs is not restricted to auditing costs but instead encompasses a much broader set of costs associated with financial distress. Bernanke and Gertler (1990) also argue that policies such as debtor bailouts, loan subsidies, and loan guarantees can transfer wealth to creditworthy entrepreneurs, thereby increasing overall efficiency and welfare. Another simplification in Bernanke and Gertler (1989) is the overlapping-generations model, in which financial contracts necessarily last only one period. Gertler (1992) demonstrates that similar quantitative results emerge when borrowers and lenders can write multi-period contracts.

A new finding is that, with multi-period relationships, the expected future profits of the borrower can partially substitute for internal financing, thereby reducing agency costs. Because an increase in the interest rate reduces the present value of expected profits, this result implies that higher interest rates worsen the agency problem.

Bernanke and Gertler (1989) generate some novel empirical predictions later tested by Bernanke et al. (1996). The first is about non-linearity. Specifically, the model predicts that in an economy with sufficient internal finance, independent and identically distributed fluctuations in current profits have little effect on investment spending, and the financial accelerator mechanism is insignificant. By contrast, fluctuations in current profits have much larger effects on investment when internal finance is already low, such as when the economy is in a deep recession. The second implication is the “flight-to-quality” phenomenon, whereby lenders reduce lending to firms that require monitoring and shift to safe alternatives when the prospective agency costs of lending increase. In this case, one should observe credit reallocations from low-net-worth to high-net-worth borrowers in downturns. Bernanke and Gertler (1989) show that at the onset of a recession, borrowers facing high agency costs (such as consumers, small firms, and firms with weak balance sheets) should receive a relatively lower share of credit extended (the flight to quality) and hence should account for a proportionally greater part of the decline in economic activity.

Kiyotaki and Moore (1997) develop an approach that is complementary to Bernanke and Gertler (1989). Specifically, their approach is based on the observation that, in practice, the debt of non-financial firms is largely backed by collateral such as land, houses, and equipment. Moreover, fluctuations in asset values can restrict the firms’ borrowing capacity. Therefore, a firm’s internal net worth is not restricted to its current profits but also includes the price of its assets. Fluctuations in asset prices can dramatically change the firm’s net worth and borrowing costs. Kiyotaki and Moore (1997) identify a feedback loop between borrowing capacity and asset prices. An initial positive shock is further amplified by an increase in asset prices, which then feeds back into more investment and output, further increases in asset prices, and so on. There is also now an added intertemporal ingredient, as asset prices respond not only to current movements in output but also to expectations of future movements.

Bernanke et al. (1999) build the financial-accelerator mechanism into a dynamic, quantitative, general equilibrium model. To this day, the model of Bernanke et al. (1999) has become the work-horse model used by policymakers and researchers to study credit markets, monetary policy, and the macroeconomy. The model takes the classic real business-cycle model and introduces credit market frictions, such as costly state verification, as well as non-credit market features, such as price stickiness, investment lags, and firm heterogeneity; but there is no bank in the model. Methodology-wise, the model is log-linearized and calibrated to study how the financial accelerator propagates the impact of monetary policy shocks. Both fronts are augmented by the recent macro-finance dynamic general stochastic equilibrium models with a banking sector, which are largely motivated by the key empirical observation that financial institutions played a central role in generating the non-linear systemic risk during the 2007–2008 global financial crisis (He and Krishnamurthy, 2013; Brunnermeier and Sannikov, 2014).

4.3. Banks and monetary policy transmission

The financial-accelerator theory has established that credit market frictions can amplify business-cycle shocks. Given that assumption, one should expect that monetary policy could be transmitted to the real economy by directly changing the magnitudes of these credit market frictions. Specifically, changes in monetary policy could affect the external finance premium, defined as the difference in cost between funds raised externally (such as raising debt or equity) and funds raised internally (such as retained earnings). Empirically, Bernanke and Blinder (1992) document that fluctuations in the Federal funds rate can forecast future real macroeconomic variables. They assume that banks cannot frictionlessly replace retail deposits with other sources of funds, such as certificates of deposits or equity. Thus, the Federal funds rate affects the supply of bank reserves. Bernanke and Blinder (1992) present evidence consistent with the view that monetary policy works through "credit" (i.e., bank loans) and through "money" (i.e., bank deposits). Bernanke and Gertler (1995) further classify the transmission of monetary policies into two channels. The balance-sheet channel emphasizes how changes in monetary policy affect borrowing firms' balance sheets and, in particular, their net worth. The second channel, the bank-lending channel, focuses on how monetary policy shocks change the supply of loans by banks. Bernanke and Gertler think a balance sheet channel seems well established, whereas the bank lending channel is more controversial due to institutional changes.[13] Later research, such as Kashyap and Stein (2000), provides more support to the bank-lending channel by showing that the impact of monetary policy on lending behavior is stronger for banks with less liquid balance sheets.

The traditional bank-lending channel works as follows. Both bank reserves and bank deposits increase following expansionary monetary policy, resulting in more loanable funds by banks. Borrowers, especially small firms, depend on bank loans to finance their activities. Therefore, investments rise and,

[13] For example, they argue that the issuance of large certificates of deposit has become easier, so the substitution for retail deposits is also more flexible.

potentially, consumer spending will also increase. For the bank lending channel to hold, there are three crucial assumptions. First, banks face reserve requirements so that demand deposits are directly related to the availability of reserves. Second, banks cannot easily substitute demand deposits with other financing methods, such as certificates of deposit or equity. Finally, firms cannot substitute bank loans with other forms of finance. Therefore, a reduction in loan supply will lead to a depression in real economic activity. Another channel through which monetary policy interacts with the banking sector is analyzed in Diamond and Rajan (2006). When the economy is subject to aggregate shocks and bank deposits (i) are promises to repay real goods (or their respective value) and (ii) cannot be made contingent on the realization of such shocks, banks might be forced to scramble for real liquidity to keep their contractual obligations with depositors. This, in turn, might reduce lending to the economy. Diamond and Rajan (2006) introduce nominal deposit contracts, that is, a promise by banks to repay depositors with money. With this contract, banks' real deposit obligation (i.e., the nominal value divided by the price level) becomes state-contingent because the price level adjusts in response to aggregate shocks, preventing credit crunches. The downside of nominal deposit contracts is that if the price level changes because of shocks to money demand – that are unrelated to real factors – the real value of banks' obligations will also be affected. This might generate bank defaults that would have been prevented with a real deposit contract. Central bank interventions that change the money supply might be needed to accommodate the shocks to money demand and prevent such shocks from generating distress in the banking system.

4.4. Banks and crises

Banks are often at center stage in financial crises. In fact, empirical research sometimes defines financial crises as episodes with banking panics and systemic bank runs. Moreover, bank failures can be contagious due to events that cause depositor panic or the interconnected contractual arrangements across banks. Diamond and Rajan (2005) identify an additional channel whereby bank failure can shrink the common liquidity pool and exacerbate the aggregate liquidity shortage. Their model builds upon Diamond and Rajan (2001) where human capital is necessary for both real production and loan collection, which creates a commitment problem. The solution to the lack of commitment problem is for the banker to issue demand deposits. They introduce aggregate shocks to production timing. However, information about aggregate liquidity can arrive even before early production. In this case, demand deposits also create a potential mismatch between the production uncertainty and depositors' demand. If the size of this aggregate mismatch is moderate, then the bank can raise additional resources against loans for delayed projects. In the general equilibrium, these additional resources must come from borrowers whose projects are early and have produced goods. However, when too many projects are delayed so that the aggregate liquidity shortage is large, then bankers will really struggle to obtain liquidity. They can attempt to raise the interest rates for new deposits, but given the aggregate shortage, the increase in interest rates cannot clear the deposit market. Instead, bankers must call loans and restructure late projects into immediate resources. Such restructuring increases the pool of available resources for liquidity demand but comes with long-run production reduction. As the deposit rates (real interest rates) increase, the bank's asset value also drops, so eventually, some banks become insolvent. This can further exacerbate liquidity shortages because depositors will withdraw immediately and demand liquidity when they anticipate future insolvency. Combined together, this leads to an illiquidity–insolvency spiral. As a result, a shortage in aggregate liquidity can result in systemic bank failures, which harms the entire banking sector.

Diamond and Rajan (2011) further argue that bank liquidity management is fundamentally inefficient. In their model, banks are subject to liquidity shocks in the future, such as an unusual increase in liquidity demand from their depositors. Banks may need to fire-sell their long-term illiquid assets to satisfy additional withdrawals. Such *future* fire sales create very profitable arbitrage opportunities for experts who know how to evaluate and deploy the assets. However, when experts anticipate future fire sales, they may also pass on some profitable lending opportunities *upfront*. One can interpret these experts as healthy banks, which are not subject to liquidity shocks. This model then implies that future fire sales by distressed banks can reduce upfront lending by healthy banks. Somewhat surprisingly, the management of the distressed bank, knowing that a liquidity shock might occur in the future and that the bank could fail, does not have incentives to sell the illiquid asset today, even though such early sales could save the bank in the future. Therefore, the bank's private liquidity management policy is fundamentally inefficient. The reason for this is that, by selling the asset today, the bank will raise cash, which bolsters the value of its depositors, but it thereby sacrifices the returns to its equity holders if the bank manages to survive – a form of risk shifting via illiquidity. Instead, the bank's management would rather hold on to the illiquid assets and risk a fire sale. In fact, if the bank had cash, it would prefer to buy more of the illiquid assets and become "illiquidity seekers". Diamond and Rajan (2011) imply that cleaning up the financial system can contribute to the recovery. They also imply that if regulators force institutions to sell illiquid assets in a timely fashion, this can enhance the overall stability of the entire banking sector.

Whereas Diamond and Rajan (2011) emphasize the benefits of liquidity, Diamond et al. (2020) show that liquidity could have downsides. In their model, there is upfront competition for assets, and experts with limited wealth borrow as much as possible (against the firm's assets) to bid enough to be successful – this is just a modeling device to ensure the corporate sector levers up to its full capacity. Lenders depend on high future bids by other outside experts to enforce debt claims. These bids are enhanced by both higher cash flow pledgeability (set by the incumbent after buying the firm) and the liquidity (that is, wealth) of possible future bidders. A sharp increase in anticipated liquidity both enhances upfront borrowing, as well as depresses the pledgeability the incumbent sets. The deterioration in pledgeability is not a problem when high liquidity is sustained. However, it becomes problematic when liquidity dries up, as there is very little supporting corporations' ability to borrow. Put differently, high expectations of liquidity create the conditions where corporations become dependent on continued future liquidity to roll over their debt. When it does not materialize, they experience a sudden stop. This can occur even if economic prospects for corporations are still bright. In these episodes, productive assets may have to be sold to others who do not know how to deploy them. As a result, economic downturns are prolonged, and recoveries are sluggish.

In Diamond et al. (2020), insiders such as firm managers choose corporate governance and pledgeability. In practice, they are also affected by the outsiders, such as financial intermediaries, through monitoring and covenants. Diamond et al. (2022) develop a theory of corporate lending by financial intermediaries under time-varying liquidity. The main results are that, starting from a low level, higher prospective corporate liquidity will initially reduce monitored borrowing from a bank in favor of arm's length borrowing; then it will steadily raise the amount corporations that can borrow at arm's length; and eventually it will reduce the need for internal corporate governance to support corporate borrowing. In parallel, higher prospective corporate liquidity will allow banks to operate with less capital or higher leverage.

5. Bank regulation policies

Bank regulation has been a controversial question for almost a century. Dewatripont et al. (2010) provide a brief introduction to the history. Roughly speaking, the US government has imposed restrictions such as deposit insurance, the deposit-rate ceiling, entry and branching barriers, capital requirements, regulator supervision, and, more recently, stress tests as well as liquidity requirements. Whereas there is still some debate around the goal of bank regulation, there seems to be a consensus (at least after the global financial crisis) that some forms of government intervention are needed to ensure the stability of banks and the banking system.

Deposit insurance was introduced in the United States in 1933. There has always been opposition to such a plan because people believe the deposit insurance system would be unduly expensive and would unfairly subsidize poorly managed banks (Federal Deposit Insurance Corporation, 1998). Before Diamond and Dybvig (1983), academic papers (Kareken and Wallace, 1978; Dothan and Williams, 1980) focused more on the moral hazard problem introduced by deposit insurance. In these models with a complete market, deposit insurance is redundant and encourages unnecessary risk-taking by banks. Diamond and Dybvig (1983) isolate the bank's choice of a risky technology and therefore show that deposit insurance provided by the government can dominate the contracts offered by banks without the insurance. Combining both insights, the choice of deposit insurance entails a tradeoff between liquidity risk-sharing and banks' incentives for risk-taking. Deposit insurance differs from the common tax and subsidy schemes, which often introduce distortion. Instead, the role of such a policy is to eliminate panic-based bank runs and prevent a bad equilibrium. A similar mechanism emerges in a contemporaneous paper (Dybvig and Spatt, 1983), albeit in a different context.

Another institutional arrangement to stop bank runs is to have a lender of last resort. The lender-of-last-resort justification of central bank lending has a long history, which goes back to Bagehot (1873). Bagehot's dictum is famously summarized by Tucker (2009) as follows: "to avert panic, central banks should lend early and freely (ie, without limit), to solvent firms, against good collateral, and at 'high rates'." In the United States, this function has been fulfilled by the discount window created in 1913. In theory, the discount window should work in a very similar way to deposit insurance. It could also induce excessive risk-taking by banks when they anticipate being bailed out. In practice, however, it is widely believed that a stigma is associated with borrowing from the discount window so that at the onset of the 2007–2008 global financial crisis, it was not much used despite the system-wide liquidity shortage, as documented in Bernanke (2015). Before the Fed, the New York Clearing House Association (NYCH), a group of 60 New York City banks, was effectively a private lender of last resort in response to banking runs in the US National Banking Era (Gorton and Tallman, 2016). The clearing house would often suspend payments during a financial crisis (Gorton, 1985), referred to as the suspension of convertibility. Diamond and Dybvig (1983) show that demand deposits and suspension of convertibility can prevent bank runs if the fraction of early types is certain. In such a situation, suspension never occurs in equilibrium. When the fraction of early types is stochastic, however, suspension can no longer provide the most efficient allocation.

Diamond and Rajan (2012) study the optimal intervention by a social planner when liquidity demand can exceed supply in some future states of the world. As established in their previous work, demand deposit solves the commitment problem, but the resulting contracts are non-state-contingent. This lack of state contingency can trigger runs, which leads to early liquidation of late projects and welfare losses. The social planner has tax authority but cannot commit to not bailing out banks if a run occurs. Like banks, the planner can neither offer state-contingent contracts nor observe the households' types being early or late. Diamond and Rajan (2012) show that a direct bailout – the planner taxes households and transfers to banks – actually makes households worse off, even though it reduces runs. This is because the disciplinary role of deposits is reduced when banks know they can be bailed out in a run. Banks can even default strategically when they are solvent, in which case the social planner is forced to intervene and offer rent to bankers. Whereas competition among banks offsets these rents, the resulting banking system becomes very levered. Consequently, the system fails in even more *ex post* states of the world. An alternative is for the planner to effectively act as an intermediary, who borrows from households and, in turn, makes loans to banks at market-determined interest rates. Diamond and Rajan (2012) show that interest rate intervention – the social planner lends to solvent banks to reduce interest rates – will dominate direct bailouts because it prevents lending to insolvent banks. They also suggest that the social planner should raise interest rates in normal times above the market-determined level to offer banks incentives to maintain low leverage and high liquidity.

Since the 2007–2008 global financial crisis, regulators around the globe have designed a set of international banking regulations to improve the banking sector's stability, safety, and resilience, known as Basel III. One of its key objectives is to improve banks' liquidity and their ability to meet their short-term obligations. Basel III established two liquidity ratios: the Liquidity Coverage Ratio (LCR) and the Net Stable Funding Ratio (NSFR). The LCR requires banks to hold sufficient high-quality liquid assets to cover expected cash outflows over a 30-day stress period. The NSFR aims to ensure that a bank's funding structure is stable over one year. By imposing these liquidity requirements, Basel III helps to ensure that banks can withstand periods of stress and continue to meet their obligations, thus reducing the risk of financial contagion and contributing to a stable financial system. Diamond and Kashyap (2016) study liquidity requirements and highlight the fundamental economic failure these requirements can solve. They study a modified version of Diamond and Dybvig (1983) but allow the bank to hold a liquid asset. Moreover, the bank has private information on the fraction of depositors who need to withdraw early for fundamental reasons. Some depositors will receive a sunspot signal about the bank, which could lead them to a run. The imperfect information creates a challenge for the banks because their customers will not necessarily know if the bank holds liquidity or not, which reduces the banks' incentives to hold liquidity. In this model, banks face a tradeoff between investing in a liquid asset that fortifies themselves against a run and forgoing profits from deterring the run. The additional liquidity to survive a run will become excessive whenever a run is avoided.

Then, why is there a need for additional liquidity? Charles Goodhart, a British economist, made a metaphor known as the last taxi problem (Goodhart, 2008). "The weary traveler who arrives at the railway station late at night, and, to his delight, sees a taxi there who could take him to his distant destination. He hails the taxi, but the taxi driver replies that he cannot take him, since local bylaws require that there must always be one taxi standing ready at the station." Unused liquidity on the bank's balance sheet is similar to the metaphorical last taxi at the station. It ensures that the bank will always have enough liquidity. This turns out to be useful because if depositors know that there will always be unusable liquidity on the bank's balance sheet, this means that the bank must also have sufficient funds to "back up" the unusable portion – the last taxi. If some depositors withdraw early, a fraction of the required liquidity can be used, but some liquidity must remain unused on the bank's balance sheet. Metaphorically, if there is always a last taxi at the station, and people know it, then there will never be a reason to panic about the existence of taxis. Diamond and Kashyap (2016) discuss the recent liquidity regulation and conclude that an LCR is better than an NSFR because it requires unused liquidity against deposits. However, the LCR from Basel III is still imperfect because, within the 30-day stress period, the bank may still use its liquidity and leave the remaining liquidity unknown to depositors. This uncertainty can still cause panic, and the fear of fear itself can cause runs.

6. Follow-up work

We briefly introduce some work that followed the contributions made by the three laureates in the early 1980s.

Coordination problems. Diamond and Dybvig (1983) do not make direct predictions on the probability of bank runs. Using the techniques from the literature of the global game, Goldstein and Pauzner (2005) derive a unique equilibrium in which runs occur if and only if the fundamentals deteriorate sufficiently.

Interbank network. Diamond and Dybvig (1983) modeled one single bank, interpreted as the entire financial intermediary industry. They mention that if many banks were introduced into the model, then there would be a role for liquidity risk-sharing between banks. Allen and Gale (2000) introduce such a model and highlight the interbank network structure. They show that contagion can occur due to overlapping claims banks have on one another.

Dynamic runs under debt contracts. Diamond and Dybvig (1983) highlight the synchronous coordination problem across depositors.

He and Xiong (2012) study the dynamics of runs and emphasize the asynchronous coordination problem among creditors who roll over debt at different times. He and Manela (2016) investigate the role of information acquisition when demand depositors face uncertainty over the originating time of a rumor that triggers other depositors to start a run on the bank. Finally, Martin et al. (2014) focus on short-term collateralized borrowing and investigate the role of the microstructure of funding markets – especially the differences between the tri-party repo market and the bilateral repo market – in driving expectations-driven runs.

Empirical evidence on bank runs. The empirical studies on bank runs are limited due to the availability of data. Using data for a bank in India that experienced a run when a neighboring bank failed, Iyer and Puri (2012) empirically test the role of deposit insurance and examine factors that affect depositors' incentives to run. Artavanis et al. (2022) use the high-frequency withdrawal data on demand and time deposits from a large Greek bank and show that about two-thirds of this increase is driven by direct exposure to deteriorating fundamentals while the remainder is due to strategic complementarities.

Liquidity mismatch in macroeconomic models. The financial-accelerator models emphasize the role of net worth in reducing agency costs. Experts in these models are sometimes interpreted as banks, but they can also be interpreted as entrepreneurs of non-financial firms. Gertler and Kiyotaki (2015) build models of liquidity mismatch – unique to banks – into the standard macroeconomic framework. They show that bank runs occur when bankers have low net worth, such as during recessions.

Fully dynamic models. In both Bernanke et al. (1999) and Kiyotaki and Moore (1997), the amplification mechanisms are studied around the deterministic steady states, and the model is solved after being log-linearized. Two recent papers, by He and Krishnamurthy (2013) and Brunnermeier and Sannikov (2014), construct fully stochastic models and focus on the global equilibrium dynamics, not just near the steady states.

7. The run on Silicon Valley Bank in March 2023

The papers cited by the Nobel Prize Committee continue to have a lasting impact on the field of study and persist until today. On Thursday, 9 March

2023, depositors withdrew $42 billion from SVB in a single day. After it failed to raise capital, SVB was closed down by California regulators on Friday, 10 March, and was placed under the receivership of the FDIC. According to the *New York Times*, the failure of SVB is the second-largest in US history and the largest since the 2007–2008 global financial crisis.[14]

At least two factors triggered the run on SVB. First, since 2021, the SVB has invested a large fraction of its portfolio in long-term bonds, particularly treasuries and mortgage-backed securities. After the Federal Reserve increased the interest rate in the second half of 2022, the market value of these bonds fell substantially. SVB clearly has insufficient hedging against interest rate risks, and it was later reported that the bank had effectively no chief risk officer for over nine months. Second, most of SVB's depositors are technology firms and startups, with deposit size way over the FDIC insurance coverage ($250,000). Right after SVB announced that it sold a portfolio of bonds of around $21 billion at a loss of $1.8 billion on Wednesday, 8 March, the uninsured start-up corporate customers panicked and rushed to withdraw their deposits the next day.

The failure of SVB quickly generated concerns for other banks. The government stepped in immediately to calm the panic, aiming to stop the contagion. The Treasury, the Federal Reserve, and the FDIC issued a joint statement on 12 March to restore confidence in the financial system. Specifically, the government guaranteed all SVB's deposits, including those uninsured. Similar measures were taken for the Signature Bank and the First Republic Bank. Finally, the Fed announced a new lending facility – a variant of the traditional lender-of-last-resort discount window operations – to provide liquidity to eligible institutions in case of excessive withdrawals; importantly, this facility allows institutions to pledge Treasury collaterals at their par value instead of market value. The bailout of uninsured deposits is not without controversy. In fact, it is likely to spark a debate on bank runs, deposit monitoring, and moral hazard.

References

Aliber, R. Z., Kindleberger, C. P., and Solow, R. M. (2015), *Manias, Panics, and Crashes: A History of Financial Crises*, Springer, Berlin.

Allen, F. and Gale, D. (2000), Financial contagion, *Journal of Political Economy 108*, 1–33.

Artavanis, N., Paravisini, D., Garcia, C. R., Seru, A., and Tsoutsoura, M. (2022), One size doesn't fit all: heterogeneous depositor compensation during periods of uncertainty, National Bureau of Economic Research Working Paper 30369.

Bagehot, W. (1873), *Lombard Street: A Description of the Money Market*, Henry S. King & Co., London.

[14] See https://www.nytimes.com/2023/03/10/business/silicon-valley-bank-stock.html.

Bernanke, B. S. (1983), Nonmonetary effects of the financial crisis in the propagation of the great depression, *American Economic Review 73* (3), 257–276.

Bernanke, B. S. (2004), Money, gold, and the great depression, Remarks by Ben S. Bernanke, Member of the Board of Governors of the US Federal Reserve System, at the H. Parker Willis Lecture in Economic Policy, Washington and Lee University, Lexington, Virginia, https://www.federalreserve.gov/boarddocs/speeches/2004/200403022/default.htm.

Bernanke, B. (2015), *The Courage to Act: A Memoir of a Crisis and its Aftermath*, W. W. Norton & Co., New York.

Bernanke, B. S. and Blinder, A. S. (1992), The federal funds rate and the channels of monetary transmission, *American Economic Review 82* (4), 901–921.

Bernanke, B. S. and Gertler, M. (1987), Banking and macroeconomic equilibrium, in W. Barnett and K. Singleton (eds), *New Approaches to Monetary Economics*, Cambridge University Press, Cambridge, 89–112.

Bernanke, B. and Gertler, M. (1989), Agency costs, net worth, and business fluctuations, *American Economic Review 79* (1), 14–31.

Bernanke, B. and Gertler, M. (1990), Financial fragility and economic performance, *Quarterly Journal of Economics 105*, 87–114.

Bernanke, B. S. and Gertler, M. (1995), Inside the black box: the credit channel of monetary policy transmission, *Journal of Economic Perspectives 9* (4), 27–48.

Bernanke, B. S. and James, H. (1991), The gold standard, deflation, and financial crisis in the great depression: an international comparison, in R. G. Hubbard (ed.), *Financial Markets and Financial Crisis*, University of Chicago Press, Chicago, IL.

Bernanke, B. S., Gertler, M., and Gilchrist, S. (1996), The flight to quality and the financial accelerator, *Review of Economics and Statistics 78*, 1–15.

Bernanke, B. S., Gertler, M., and Gilchrist, S. (1999), The financial accelerator in a quantitative business cycle framework, in J. B. Taylor and M. Woodford (eds), *Handbook of Macroeconomics*, Vol. 1, North-Holland, Amsterdam, 1341–1393.

Brunnermeier, M. K. and Oehmke, M. (2013), The maturity rat race, *Journal of Finance 68*, 483–521.

Brunnermeier, M. K. and Sannikov, Y. (2014), A macroeconomic model with a financial sector, *American Economic Review 104* (2), 379–421.

Bryant, J. (1980), A model of reserves, bank runs, and deposit insurance, *Journal of Banking & Finance 4*, 335–344.

Bryant, J. and Wallace, N. (1980), Open-market operations in a model of regulated, insured intermediaries, *Journal of Political Economy 88*, 146–173.

DeMarzo, P. M. (2005), The pooling and tranching of securities: a model of informed intermediation, *Review of Financial Studies 18*, 1–35.

Dewatripont, M., Rochet, J.-C., and Tirole, J. (2010), Balancing the banks, in *Balancing the Banks*, Princeton University Press, Princeton, NJ.

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

Diamond, D. W. (1989), Reputation acquisition in debt markets, *Journal of Political Economy 97*, 828–862.

Diamond, D. W. (1991a), Debt maturity structure and liquidity risk, *Quarterly Journal of Economics 106*, 709–737.

Diamond, D. W. (1991b), Monitoring and reputation: the choice between bank loans and directly placed debt, *Journal of Political Economy 99*, 689–721.

Diamond, D. W. (1993), Seniority and maturity of debt contracts, *Journal of Financial Economics 33*, 341–368.

Diamond, D. W. (1997) Liquidity, banks, and markets, *Journal of Political Economy 105*, 928–956.

Diamond, D. W. (2004), Presidential address, committing to commit: short-term debt when enforcement is costly, *Journal of Finance 59*, 1447–1479.

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

Diamond, D. W. and Dybvig, P. H. (1986), Banking theory, deposit insurance, and bank regulation, *Journal of Business 59*, 55–68.

Diamond, D. W. and He, Z. (2014), A theory of debt maturity: the long and short of debt overhang, *Journal of Finance 69*, 719–762.

Diamond, D. W. and Kashyap, A. K. (2016), Liquidity requirements, liquidity choice, and financial stability, in J. B. Taylor and H. Uhlig (eds), *Handbook of Macroeconomics*, Vol. 2, Elsevier, Amsterdam, 2263–2303.

Diamond, D. W. and Rajan, R. G. (2000), A theory of bank capital, *Journal of Finance 55*, 2431–2465.

Diamond, D. W. and Rajan, R. G. (2001) Liquidity risk, liquidity creation, and financial fragility: A theory of banking, *Journal of Political Economy 109*, 287–327.

Diamond, D. W. and Rajan, R. G. (2005), Liquidity shortages and banking crises, *Journal of Finance 60*, 615–647.

Diamond, D. W. and Rajan, R. G. (2006), Money in a theory of banking, *American Economic Review 96* (1), 30–53.

Diamond, D. W. and Rajan, R. G. (2011), Fear of fire sales, illiquidity seeking, and credit freezes, *Quarterly Journal of Economics 126*, 557–591.

Diamond, D. W. and Rajan, R. G. (2012), Illiquid banks, financial stability, and interest rate policy, *Journal of Political Economy 120*, 552–591.

Diamond, D. W. and Verrecchia, R. E. (1991), Disclosure, liquidity, and the cost of capital, *Journal of Finance 46*, 1325–1359.

Diamond, D. W., Hu, Y., and Rajan, R. G. (2020), Pledgeability, industry liquidity, and financing cycles, *Journal of Finance 75*, 419–461.

Diamond, D. W., Hu, Y., and Rajan, R. G. (2022), Liquidity, pledgeability, and the nature of lending, *Journal of Financial Economics 143*, 1275–1294.

Dothan, U. and Williams, J. (1980), Banks, bankruptcy, and public regulation, *Journal of Banking & Finance 4*, 65–87.

Dybvig, P. H. and Spatt, C. S. (1983), Adoption externalities as public goods, *Journal of Public Economics 20*, 231–247.

Eichengreen, B. and Sachs, J. (1985), Exchange rates and economic recovery in the 1930s, *Journal of Economic History 45*, 925–946.

Federal Deposit Insurance Corporation (1998), *A Brief History of Deposit Insurance in the United States*, prepared for the International Conference on Deposit Insurance Washington, DC, FDIC Division of Research and Statistics.

Flood, R. P. and Garber, P. M. (1981), A systematic banking collapse in a perfect foresight world, National Bureau of Economic Research Working Paper 0691.

Freixas, X. and Rochet, J.-C. (2008), *Microeconomics of Banking*, MIT Press, Cambridge, MA.

Friedman, M. and Schwartz, A. J. (1963), *A Monetary History of the United States, 1867–1960*, Princeton University Press, Princeton, NJ.

Gertler, M. (1992), Financial capacity and output fluctuations in an economy with multi-period financial relationships, *Review of Economic Studies 59*, 455–472.

Gertler, M. and Kiyotaki, N. (2015), Banking, liquidity, and bank runs in an infinite horizon economy, *American Economic Review 105* (7), 2011–2043.

Goldstein, I. and Pauzner, A. (2005), Demand–deposit contracts and the probability of bank runs, *Journal of Finance 60*, 1293–1327.

Goodhart, C. (2008), Liquidity risk management, *Banque de France Financial Stability Review 11*, 39–44.

Gorton, G. (1985), Clearinghouses and the origin of central banking in the united states, *Journal of Economic History 45*, 277–283.

Gorton, G. and Tallman, E. W. (2016), Too big to fail before the Fed, *American Economic Review 106* (5), 528–532.

Gorton, G. B. and Tallman, E. W. (2018), Chapter 1. Fighting financial crises: learning from the past, in *Fighting Financial Crises: Learning from the Past*, University of Chicago Press, Chicago, IL, 1–11.

Gorton, G. and Winton, A. (2003), Financial intermediation, in G. M. Constantinides, M. Harris, and R. M. Stulz (eds), *Handbook of the Economics of Finance*, Vol. 1, Elsevier, Amsterdam, 431–552.

Hammond, B. (1991), *Banks and Politics in America from the Revolution to the Civil War*, Princeton University Press, Priceton, NJ.

Hart, O. and Moore, J. (1994), A theory of debt based on the inalienability of human capital, *Quarterly Journal of Economics 109*, 841–879.

He, Z. and Krishnamurthy, A. (2013), Intermediary asset pricing, *American Economic Review 103* (2), 732–770.

He, Z. and Manela, A. (2016), Information acquisition in rumor-based bank runs, *Journal of Finance 71*, 1113–1158.

He, Z. and Xiong, W. (2012), Dynamic debt runs, *Review of Financial Studies 25*, 1799–1843.

Hu, Y. and Varas, F. (2021), A theory of zombie lending, *Journal of Finance 76*, 1813–1867.

Hu, Y., Varas, F., and Ying, C. (2021), Debt maturity management, Working Paper.

Humphrey, D. B. (1986), Payments finality and risk of settlement failure, in A. S. Saunders and L. J. White (eds), *Technology and Regulation of Financial Markets: Securities, Futures and Banking*.

Iyer, R. and Puri, M. (2012), Understanding bank runs: the importance of depositor–bank relationships and networks, *American Economic Review 102* (2), 1414–1445.

Jacklin, C. J. (1987), Demand deposits, trading restrictions, and risk sharing, in E. C. Prescott and N. Wallace (Eds), *Contractual Arrangements for Intertemporal Trade*, University of Minnesota Press, Minneapolis, MN, 26–47.

Jacklin, C. J. and Bhattacharya, S. (1988), Distinguishing panics and information-based bank runs: welfare and policy implications, *Journal of Political Economy 96*, 568–592.

Jensen, M. C. (1988), Takeovers: their causes and consequences, *Journal of Economic Perspectives 2* (1), 21–48.

Jensen, M. C. and Meckling, W. H. (1976), Theory of the firm: managerial behavior, agency costs and ownership structure, *Journal of Financial Economics 3*, 305–360.

Kareken, J. H. and Wallace, N. (1978), Deposit insurance and bank regulation: a partial-equilibrium exposition, *Journal of Business 51*, 413–438.

Kashyap, A. K. and Stein, J. C. (2000), What do a million observations on banks say about the transmission of monetary policy?, *American Economic Review 90* (3), 407–428.

Kiyotaki, N. and Moore, J. (1997), Credit cycles, *Journal of Political Economy 105*, 211–248.

Martin, A., Skeie, D., and von Thadden, E.-L. (2014), Repo runs, *Review of Financial Studies 27*, 957–989.

Modigliani, F. and Miller, M. H. (1958), The cost of capital, corporation finance and the theory of investment, *American Economic Review 48* (3), 261–297.

Myers, S. C. (1977), Determinants of corporate borrowing, *Journal of Financial Economics 5*, 147–175.

Patinkin, D. (1965), *Money, Interest, and Prices: An Integration of Monetary and Value Theory*, MIT Press, Cambridge, MA.

Samuelson, P. A. (1958), An exact consumption-loan model of interest with or without the social contrivance of money, *Journal of Political Economy 66*, 467–482.

Sowerbutts, R., Schneebalg, M., and Hubert, F. (2016), The demise of overend gurney, *Bank of England Quarterly Bulletin*, *Q2*.

Tirole, J. (2010), *The Theory of Corporate Finance*, Princeton University Press, Princeton, NJ.

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

Tucker, P. (2009), The repertoire of official sector interventions in the financial system: last resort lending, market-making, and capital, Speech delivered to the Bank of Japan 2009 International Conference on Financial System and Monetary Policy: Implementation, Bank of Japan, 27–28 May, https://www.bankofengland.co.uk/speech/2009/last-resort-lending-market-making-and-capital.