The Regulation of Private Money*

Financial crises are bank runs. At root, the problem is short-term debt (private money), which while an essential feature of market economies, is inherently vulnerable to runs in all its forms (not just demand deposits). Bank regulation aims at preventing bank runs. History shows two approaches to bank regulation: the use of high-quality collateral to back banks’ short-term debt and government insurance for the short-term debt. Also, explicit or implicit limitations on entry into banking can create charter value (an intangible asset) that is lost if the bank fails. This can create an incentive for the bank to abide by the regulations and not take too much risk.

JEL codes: E4, G2
Keywords: Bank regulation, privatge money

Financial crises (bank runs) are the norm in market economies. They have plagued market economies throughout history, in developed and developing countries, in countries with and without a central bank, in countries with and without deposit insurance. The Panic of 2007–8 is a reminder of this fact. The goal of bank regulation is to reduce the likelihood of a financial crisis, a bank run.

Banks have always been regulated because their product, their output, is short-term debt and such debt is subject to systemic runs in which a country’s financial system can collapse without the intervention of private bank clearinghouses or the government (and central bank). But, while banks have always been regulated (either privately or publicly, or both), historically such regulation has never been successful in eliminating financial crises. At best, regulation has produced long periods without a financial crisis.

*Thanks to Bob DeYoung for comments and suggestions.

Gary Gorton is at Yale School of Management and NBER (E-mail: gary.gorton@yale.edu).

Received May 17, 2019; and accepted in revised form May 17, 2019.

1. A bank is a firm that issues short-term debt, whether it is recognized by the government as a “bank” or not. A bank clearinghouse is a private coalition or club of banks that meets to “clear” checks, that is the banks pay daily balances due to and from each other at one time and place. The clearinghouse also adopted rules and regulations to monitor banks because of the counterparty risk faced in the clearing process. During a financial crisis, the clearinghouse played an important role in mitigating the effects of the crisis (see Gorton and Tallman 2018).
What is the problem with short-term debt? It is difficult to engage in transactions or store value when the price of a claim fluctuates and parties are differentially informed about the value of the claim. This has been a problem for centuries. Even the value of gold coins was difficult to ascertain because coins were shaved or sweatied (shaken to remove gold dust) (see Gorton 2017). Gorton and Pennacchi (1990) argue that short-term debt should be designed to be immune to such problems. Users of short-term debt should not have to worry about losing to insiders when transacting. Dang, Gorton, and Holmström (2017) argue that this can be accomplished if the price of the debt does not change. And this is best accomplished by backing the debt with debt. For example, backing private bank notes with state bonds, or backing demand deposits with portfolios of loans.

The optimal contract of debt-on-debt suggests two approaches to regulating private money to maintain the price at a one-to-one ratio with cash. History shows two approaches to bank regulation: the use of high-quality collateral to back bank short-term debt and insurance for the short-term debt. But, in any case, the financial system, like other industries, changes and forms of short-term debt change. Repo, for example, grew into a large category of short-term debt over the last 40 or 50 years. When the short-term debt, and its collateral, are not regulated, it can happen that the fixed price cannot be maintained. In that case, the quantities adjust, to zero in a bank run.

While some countries have gone for long periods of time without experiencing a financial crisis, it is not clear what the banking systems of these countries have in common. Why did the U.S. not experience a financial crisis during 1934–2007? What forms of bank regulation work, in the sense of creating a long period without a crisis? And, even if the regulation works, will it work for all forms of short-term debt? What is the social trade-off between repressive bank regulation and economic growth? What is “optimal” bank regulation? We do not know the answers to these questions.

One important feature of bank regulation is the explicit or implicit limitations on entry into banking. Limited entry creates monopoly rents which results in (bank) charter value, an intangible asset that is lost if the bank fails. This can create an incentive for the bank to abide by the regulations and not take too much risk. But, limitations on entry require identifying “banks” and private bank money. History shows that this has been very difficult to do. Without an incentive to abide by bank regulations, the banking system can migrate out of the regulated sector. In other words, bank regulators essentially only determine where the banking system is: in the regulated sector or in the shadows.

In this essay, I discuss some of what we do and do not know about bank regulation.

1. FINANCIAL CRISES

The output of banks is short-term debt, private money (see Diamond and Dybvig 1983, Gorton and Pennacchi 1990). General Motors produces cars. McKinsey produces advice. Banks produce short-term debt. There are many kinds of short-term debt that banks issue today or have issued in the past (and will issue in the
future). These include private bank notes, bills of exchange, demand deposits, sale and repurchase agreements, various forms of commercial paper, and money market mutual funds. While these various forms of debt differ in important ways, all these forms of short-term debt are subject to runs.

While countries have had different experiences with respect to financial crises, it is clear that financial crises plague all market economies and have throughout history. This point has been made by, for example, Kindleberger (1978, 1993), Reinhart and Rogoff (2009), and Cassis (2011)—crises occur over and over. Reinhart and Rogoff (2013) show that the incidence of crises is the same in high- and middle-income countries. Laeven and Valencia (2012) count 147 banking crises over the period 1970–2011. And, in particular, developed economies have crises. Laeven and Valencia (2008) report that 62% of the crises in their modern era sample had bank runs. Reinhart and Rogoff (2013, p. 4559) note that “for the advanced economies during the full sample, the picture that emerges is one of serial banking crises.” Still some countries have more frequent crises than others. Some countries go for long periods with no crisis. Nevertheless, it is clear that financial crises plague all market economies.

The basic point that bank runs occur in all market economies has been obscured by modern crises that occur in the presence of a central bank. In that case, agents expect the central bank to act when there is the possibility of a run. Agents wait. But, the intervention of the central bank is late and the run comes even later in the midst of chaotic events. The run may be latent or quiet (i.e., deposits are withdrawn slowly but in significant amounts). Modern crises’ start dates chosen by researchers are based on observed events, such as the announcement of a blanket guarantee of bank debt, bank nationalization, or a bailout. But these are in response to some underlying events, which outside observers do not see. Boyd, De Nicolò, and Rodionova (2019) study the four leading crisis classifications and dating of modern crisis events. Boyd et al. show that extreme drops in loans can predict modern crisis start dates and that the true start date is prior to the date listed in crisis databases. In other words, there was something going on that was known by banks, which cut lending. Depositors appear to wait to see what the government will do, but banks are proactive and reduce lending.

The chaos of modern crises obscures the underlying cause. The issue is important. The problem of understanding financial crises in the presence of a central bank is most evident with respect to the Great Depression, an event which has plagued macroeconomics. At the outset of modern macroeconomics, explaining the Great Depression

---

2. In the remaining cases, there was usually the issuance of a blanket guarantee, bailouts, nationalization or other intervention that calmed debt holders.

3. One important factor in determining the frequency of crises is the industrial organization of the banking system. See, for example, Calomiris and Gorton (1991).

4. Or, there is no run, often because of a blanket guarantee by the government, nationalization of banks, or bailouts.

5. See Gorton and Tallman (2018) for some specific examples.

was viewed as an insurmountable barrier. Lucas (1980): “… the [Great] Depression continues … to defy explanation by existing economic analysis” (p. 706). And, Prescott (1983): “How can a theory claim to explain the business cycle if it cannot explain the Great Depression” (p. 11); “The answer to [this] question is simply that competitive equilibrium theory is not suited to modelling economic fluctuations in periods of great political and financial institution instability” (p. 12), that is, a financial crisis. While subsequent researchers did pick over the carcass of the Great Depression, this work did not address the central question of the role of financial crises in business cycles generally. Largely this was because the history of crises prior to the Great Depression were not studied, making the Great Depression seem extremely unusual (see Gorton, Laarits, and Muir 2019).

The crisis of 2007–8, however, was a classic bank run. There was a run on repo at the beginning (see Gorton and Metrick 2012, and Gorton, Laarits, and Metrick 2018). But, because the run happened on trading floors via Bloomberg messaging and phone calls, it was not observed by outsiders, who then only saw the effects of the run. And the effects were then interpreted as causes. Being unfamiliar with other financial crises, and, because the underlying cause of the 2007–8 crisis was not understood, the cause of the 2007–8 crisis is taken to be the large failures of banks due to greed and incompetence and prospect of bailouts (moral hazard). Never mind that these ideas suggest that there should have been a crisis sometime during the period 1934–2007. As a result, crises are not seen as at root being about short-term debt, nor have crises been seen as an inherent part of macro-economic activity. This confusion reappears when it comes to designing bank regulation.

2. SHORT-TERM DEBT

Short-term debt is an inherent feature of market economies, like private property rights, transactions occurring through markets with prices determined by demand and supply, free competition, etc. What is the problem with short-term debt? Short-term bank debt is needed for transactions and to store value over short periods of time. To be effective for these purposes, the debt needs to trade at par and the face value must be preserved with near certainty over short periods. The face value must be viewed as unquestioned. In other words, the price should be constant.

7. Historically, when observers saw depositors or note holders literally lined up at banks to withdraw, it was still not clear what caused this. So, seeing the event has not been sufficient to produce an understanding of it.

8. Financial crises are part of movements in aggregate economic activity. Crises do not happen at random times, but typically occur following credit booms and near the peak of the business cycle. Gorton (1988) studied the U.S. National Banking Era (1864–1914), a period during which banking panics regularly occurred, and he shows that the arrival of news forecasting a recession that was sufficiently bad arrived near business cycle peaks. Dimsdale and Hotson (2014) summarize the UK experience since 1825: “The general pattern is one in which financial crises occur close to business cycle peaks, and are followed by a downturn in the wider economy” (p. 26). In the modern era, the links between financial crises and recessions are similar. For example, Demirgüç-Kunt and Detragiache (1988) study a large sample of developed and developing countries over the period 1980–94 and find that “crises tend to erupt when the macroeconomic environment is weak” (p. 81); see also, for example, Kaminsky and Reinhart (1999).
It is easiest to see this by looking at the U.S. Free Banking Era, a time in which banks issued their own private currency. Depending on the exact year, there were about 1,500 currencies circulating at various discounts to the face value (when some distance from the issuing bank). These discounts varied over time (see Gorton 1996, 1999). But, varying discounts made actual transactions (and legal contracting) very difficult. There was constant haggling and arguing over the value of notes in transactions. Private bank notes were not an economically efficient means of payment. As Ricardo put it:

In the use of money, everyone is a trader; those whose habits and pursuits are little suited to explore the mechanism of trade are obliged to make use of money, and are no way qualified to ascertain the solidity of different banks whose paper is in circulation; accordingly we find that men living on limited incomes, women, laborers, and mechanics of all descriptions, are often severe sufferers by the failure of country banks … Ricardo (1876, p. 409)

There were a number of problems with a private money where the price of the money changed. The bank note market (for secondary trading) was efficient. A bank note is a perpetual zero coupon bond with an embedded put option allowing the holder to ask for par in gold or silver coin at the issuing bank. The maturity of the put option is the time it takes to return from the note holders place to the bank. For example, an agent in Philadelphia holding a note issued by a New Orleans bank would have to go back to New Orleans to exercise the put option. Gorton (1999) calculates the implied volatilities from these put options and shows in panel data that the note market was efficient in the Fama sense. But, although the note market was efficient in that sense it was not economically efficient. It was difficult to transact because of constant disputes about the correct discounts. It was also difficult to write contracts.

The secondary market note discounts were efficient because of trading by bill brokers. These were the “informed traders.” Appleton (1831):

This state of [circulating private bank notes] introduced a new branch of business and a new set of men, that of money brokers, whose business it was to exchange these currencies, one for the other, reserving to themselves a commission of about $\frac{1}{4}$ of one per cent.

The state of the currency became the subject of general complaint, the brokers were denounced as the authors of mischief… (p. 11)

The solution? Ricardo (1816): “All writers on the subject of money have agreed that uniformity in the value of the circulating medium is an object greatly to be desired” (p. 7). The problem is how to achieve this (see Gorton 2017).

The idea that the value of bank debt should not change (so as to avoid adverse selection when transacting) was put forth by Gorton and Pennacchi (1990). They argued that banks issuing debt were optimal institutions that produced private money that protected uninformed traders from losses to informed traders. They did not
show that debt was an optimal contract. Dang, Gorton, and Holmström (2017) (DGH) show that debt is the optimal security for trading because it minimizes the incentive for a counterparty to produce private information about the payoff on the trading security. In other words, transactions are efficient if the debt offered as payment is simply unquestioningly accepted (i.e., the discount is zero). No agent has an incentive to produce private information about the backing collateral; it is too expensive. And everyone knows this. DGH also show—and this is the main result—that the optimal collateral for the backing for the debt is also debt. Debt-on-debt, that is debt backed by debt, is the optimal contract. The reason is that debt-on-debt maximally makes debt information-insensitive.

Indeed, all banks’ short-term debt is backed by debt. Free bank notes were required to be backed by state bonds. Demand deposits are backed by portfolios of loans. Repo is backed by specific bonds. Asset-backed commercial paper is backed by asset-backed and mortgage-backed securities. And so on. Notably, the backing collateral itself is information-insensitive so that it is accepted as collateral without due diligence. How else would overnight repo, or 3-day commercial paper, work, for example, if lenders wanted to conduct due diligence on the backing collateral.

Debt-on-debt is the structure which optimally makes the debt worth par, Ricardo’s uniform value. Debt-on-debt is optimally information-insensitive, which means it is liquid.10 “Information-insensitive” means that it is common knowledge that no agent finds it optimal to produce (costly) private information about the fundamentals behind the debt.11 In other words, the price of this debt should not change. Private money is the one case where we do not want the price system to work. But, if prices do not change, the only thing that can adjust is quantities. If holders of the short-term debt come to suspect the backing collateral then the quantities may adjust to zero in a run. Because no one knows how to value this collateral, there is no market for it. See Hanson and Sunderam (2013). So, for the lenders the risk is that the collateral, and hence the short-term debt, become information-sensitive, in which case no one knows the value of the collateral. This is a risk that cannot be overcome by, for example, increasing the interest rate paid on overnight repo. If the collateral becomes information-sensitive, no one knows what appropriate interest rate should be, which is why quantities adjust.

Not only is it optimal to back the short-term debt with debt, but in order for a bank to produce information-insensitive liabilities its assets, for example, its loan portfolio, needs to be hard to value and details about the assets need to be kept secret (see Dang et al. 2017). Historically, this opacity developed endogenously once bank notes were replaced by demand deposits (see Gorton 2015).12 In other words, banks are optimally

10. Notably, the other case of liquidity is information-sensitive as with equity. Thus, there are two kinds of liquidity: safe liquidity (information-insensitive) and risky liquidity (information-sensitive) (see Holmström 2015).

11. Holmström (2015) draws a number of other important implications from these ideas. Also, see Dang, Gorton, and Holmström (Forthcoming).

12. The National Bank Acts, 1863–65, established national bank notes and subsequently private bank notes were taxed out of existence.
opaque. In fact, secrecy surrounds banking. For example, bank examination is two centuries old, but examination results are always confidential. In New York, it was a crime to spread false rumors about banks (see, e.g., N.Y. Banking Law § 671 False Statements or Rumors as to Banking Institutions). A final example, from the Report of the Committee of the House of Commons, appointed to enquire into the operation of the Act of 7th Geo. IV. Chapter 46 (1836):

The [bank] directors are in general required to sign a declaration, pledging themselves to observe secrecy as to the transactions of the bank with their customers, and the state of the accounts of individuals. In some of the companies [banks] this declaration has to be signed by all the clerks and officers. One banking company goes so far as to require an oath to this effect. (Quoted in Gordon 1836, p. 43)

The problem is that if information arrives that (rationally) causes the debt holders to suspect the backing collateral (it could be a portfolio of loans, for example), then the debt can switch from information-insensitive to information-sensitive—a crisis.13 This switch is the loss of confidence. It has been understood for quite a while. For example, English judge Pemberton Leigh, Chancellor of the Duchy of Cornwall, in The Bank of Australasia v. Breillant, 6 Moore’s P.C., 152 (1847), 194:

The nature of a banker’s business, especially if the bank be one both of issue and deposit, necessarily exposes him to sudden and immediate demands, which may be to the extent of a large portion of his debts, while his profits are to be made in discounting bills, in loans and other modes of investment. It is impossible that he should always have his assets in such a state as to be applicable immediately to the payment of all demands that may be made upon him… “ (also quoted in Curtis v. Leavitt 15 N.Y. 9; p. 52 (1857)).

So, banks are regulated because a systemic run can bring down the banking system, even though the banking system as a whole is fundamentally solvent.

3. BANK REGULATION

What should be done about the vulnerability of bank debt to runs? The debt-on-debt paradigm suggests that either the backing debt (the collateral) should be required to be high-quality so that its value is (almost) never questioned by the holders of the short-term debt. Or, the short-term debt should be insured, in which case, if the insurance is credible, the backing collateral becomes irrelevant. Indeed, these are the two fundamental approaches to bank regulation that we have seen since the late 1700s and which continues today.

13. Benemelech and Bergman (2017a, 2017b) test a prediction of the Dang, Gorton, and Holmström (2017), namely, that when the likelihood of bond default is low there is no incentive to produce private information about the bond’s fundamentals so the bond will be information-insensitive, there is, liquid; but, when default risk rises and the bond’s price falls, the bond becomes information-sensitive as private information is produced. Hence, in the latter case bond illiquidity rises. In other words, bond illiquidity rises as bond value declines in a specific nonlinear way. Benemelech and Bergman (2017a) study the Panic of 1873 and the Financial Crisis of 2007–8. Benemelech and Bergman (2017b) studies the issue more generally. Their evidence is consistent with Dang, Gorton, and Holmström (2012). Also see Dang, Gorton, and Holmström (Forthcoming).
Of these two approaches, focus on the backing collateral predominated historically. The idea that bank short-term debt should be backed by high-quality debt is an old one. And, the Basel liquidity coverage ratio is the latest version of this.14

3.1 Regulation of Backing Collateral (The Collateral Approach)

In late 18th century England, this idea was often brought up because of the “disadvantage of having a security which experienced sharp fluctuations in value” (Pressnell 1956, p. 504). In 1817, a measure was introduced in the House of Commons to require that small private bank notes be backed by government securities. But, this legislation died because of the opposing view that it would drive banks out of business (see Pressnell 1956, pp. 505–6). This issue of whether banks will exit in the face of regulations will reappear below.

In England, Peel’s Act of 1844 (The Bank Charter Act of 1844) limited the amount of notes that the Bank of England could issue that were not backed by gold bullion. See Torrens (1858) and Withers (1910). This limitation was opposed on the grounds that the Bank of England would not be able to fight panics. With the limitation of note issuance, the Bank could not buy securities and introduce more notes into circulation. And, indeed, when panics occurred in 1847, 1857, and 1866, the act was suspended by parliament. The Secret Committee of the House of Lords on Commercial Distress 1847–48, concluded that “… the Act is defective and that, in consequence of these defects it aggravated the distress of 1847” (quoted in Withers 1910, p. 232). Furthermore, one view is that the restriction led to an increase in the use of demand deposits (e.g., see Feavearyear 1963, and von Mises 1936)—the rise of a shadow banking system.

Even before the Act of 1844 in England, some U.S. states tried the collateral approach to make private bank notes safe. The U.S. Free Banking Era was a period in which a number of states allowed free entry into banking under the restriction that note issuance be backed by depositing state bonds with the state treasurer.15 The exact bond requirements varied by state and over time. These requirements did not eliminate the discount on notes that circulated away from the issuing bank. And, banking panics in which note holders demanded specie (gold and silver coins) were not eliminated. The problem was that the state bonds varied in value and many states defaulted on their bonds in the 1840s. Rolnick and Weber (1984) show that free banks tended to fail when the prices of the bonds backing their notes, and held by the state treasurer, fell. There was simply a shortage of safe collateral.

In the United States, the National Banking Acts (1863, 1865) were passed to finance the North in the Civil War. The acts established national banks which could

14. The liquidity coverage ratio requires that net short-term debt be backed one-for-one with “high-quality liquid assets.” So, essentially repo must be backed by U.S. Treasuries, for example.

15. The first free banking act was passed in Michigan in 1837, but it was New York’s free banking act of 1838 which got the most attention and was copied by a number of other states. In all 18 states adopted some form of free banking.
issue their own national bank notes, but these had to be backed by U.S. Treasury bonds deposited with the U.S. Treasury. The idea was that this would create a demand for U.S. Treasuries. Private bank notes were eliminated by enacting a very high tax. It was thought that banking panics would be eliminated under the national bank note requirements, but this did not happen. Instead, the panics were incidents where holders of demand deposits would run on their banks and demand national bank notes. In fact, national bank notes were not issued to the full extent that they could have been, and demand deposits grew very fast (seven times faster in the United States than in other developed economies). See Gorton, Laarits, and Muir (2020). The growth of demand deposits was the shadow banking system of its time. As in England in the debates between the Banking School and the Currency School of thought, it was disputed as to whether demand deposits were actually money.

Overall, the attempts at creating safe money by requiring sufficiently safe backing collateral did not work. The problem was either that there was no safe debt or not enough safe debt. The scarcity of safe debt was a problem leading up the Panic of 2007–8 (see Gorton, Laarits, and Muir 2020). The problem is that high-quality collateral, like U.S. Treasuries, has a convenience yield (Krishnamurthy and Vissing-Jorgensen 2012, 2015). High-quality securities pay off at par in the future with near certainty, so they are effective for storing value. Essentially Treasuries are a form of money. As a result, there are many demands for Treasuries, for example, for use as collateral in derivative transactions, and as collateral for repo. In fact, half of U.S. Treasury debt and agency debt is held abroad. Securitization was the response. Securitization is the private production of near-safe debt.16

Currently, there is a shortage of safe debt (see Gorton and Laarits 2018). So the liquidity coverage ratio of Basel which reintroduces the backing collateral requirement will likely have the same problems as occurred during the National Banking Era, namely, a shadow banking system will develop (see Gorton, Laarits, and Muir 2020).

Another way to improve the backing collateral is leverage restrictions or, equivalently, capital requirements. For example, New York in “An Act to incorporate the Stockholders of the Farmers’ and Mechanics’ Bank in the City of Schenectady,” includes the following requirement:

*And be it further enacted, That the total amount of debts at any time and in any manner due by the said corporation over and above the specie then actually deposited in the bank, shall not exceed three times the sum of the capital stock subscribed and actually paid into the said bank;* p. 229. *Journal of the Senate of the State of New York, 36th session, 1813* (S. Southwick)

In the United States, a system of double liability of shareholders of national banks was part of the National Bank Acts of 1863, 1865. Double liability meant that a bank’s shareholders were responsible for the bank’s debt up to the par value of their shares. Some states had the system before the Civil War, as well. National banks were subject

16. Xie (2012) reports that 85% of a public securitization transaction consist of AAA/Aaa debt.
to double liability, but state banks were not uniformly faced with double liability. Some states adopted the system for their banks, while others did not. This system was abandoned with the advent of deposit insurance in 1934. One issue with double liability was that state banks under double liability operated with significantly lower capital ratios (18.2%) compared to banks in states without double liability systems (22.9%) (see Macey and Miller 1992).

But, specific capital requirements or leverage restrictions are recent arrivals. Explicit capital requirements started in the early 1980s. Prior to the 1980s, capital requirements were one of the many things that bank examiners in the United States looked at using the CAMELS rating system. This was under the regime of deposit insurance. There are several problems with capital requirements which explain why historically an explicit rule was never by itself a pillar of bank regulation. The main problem is that the question of how much capital should be required is very hard to answer. Indeed, we have no good method. It was easier to require one-to-one backing with high-quality assets, although this approach has generally failed. Another issue is that in general equilibrium the equity must come from somewhere. Either depositors must be convinced to hold equity (in which case they lose the convenience of bank debt) or equity holders in the nonfinancial sector must switch to holding bank equity, in which case the banks must make up the difference with loans, so the loan book is now riskier. There is no easy solution (see Gorton and Winton 2017).

3.2 Bank Note and Deposit Insurance (The Insurance Approach)

In the United States and in Continental Europe, there has for many decades been a tradition of deposit insurance—bank deposits are guaranteed by the state up to some limit. By contrast, Canada adopted deposit insurance only in 1967 and until 1979 the United Kingdom had no deposit insurance whatever (it was introduced in 1979 only as the result of a European Directive) and until 2007 only the first £2,000 of deposits were fully insured.

The Safety Fund statute of the State of New York passed in 1829 was the first bank deposit insurance law ever enacted. The fund insured chartered banks in New York. It was funded by limited annual contributions from member banks and regulated by the state government. The statute included provisions for regular examinations of the banks by government officials. Severe losses in the panic of 1837 and the aftermath made it unable to repay losses and it ended. It lasted until 1842 when the fund was insolvent. The act was then amended to only cover bank notes. It then limped along until the panic of 1857 (see Chaddock 1910).

18. Gorton and Metrick (2010) discuss the collateral approach for addressing the vulnerability of the shadow banking system.
Five states adopted insurance in the pre-Civil War era (Vermont, Michigan, Indiana, Ohio, and Iowa). Basically, insurance systems that relied heavily on self-regulation were successful when there was mutual liability (Indiana, Ohio, and Iowa).\(^{20}\) Indiana, for example, had unlimited liability for member banks but also gave the fund directors the ability to enforce laws and regulations and also to determine when to close a bank. Ohio and Iowa were similar. The compulsory state systems were not successful (see Golembe and Warburton 1958, Golembe 1960, and Calomiris 1990).

Following the Panic of 1907, eight states adopted deposit insurance (Oklahoma, Texas, Kansas, Nebraska, South Dakota, North Dakota, Washington, and Mississippi). According to Calomiris (1990), these states failed to learn the lessons from the pre-Civil War experiences. All of these funds collapsed during the recession of the early 1920s. Also, see Robb (1921). Thus, going into the Great Depression, the experience with state-run deposit insurance schemes was negative. Nevertheless there was widespread populist support for deposit insurance during the Great Depression. Basically, the American public wanted to be rid of banking panics once and for all. The debate over deposit insurance was fierce and we need not recapitulate the arguments here (see Flood 1992).

The important point about the deposit insurance scheme adopted in the U.S. Banking Act of 1933 (commonly known as the Glass-Steagall Act) was that it was successful. Banking panics with respect to demand deposits were stopped.\(^{21}\) It is important to note that the Act contained a number of other important provisions. For example, there are provisions that sought to prohibit “speculative” credit and Regulation Q which limited the level of interest that could be paid to depositors.

Does deposit insurance work more generally, around the world? Demirgüç-Kunt and Detragiache (2002, 2005), Barth, Caprio, and Levine (2004, 2006), Beck et al. (2006), and Anginer et al. (2014) study this question in international panels. They find that the probability of a banking crisis is significantly higher in countries with an explicit deposit insurance system in place. These findings have been interpreted as consistent with the moral hazard incentives created by deposit insurance.

However, these studies date the start of a crisis using the “events method,” that is the start is determined by the time of observed interventions by the central bank or government. But, by then the central bank or government is responding to some underlying events. As discussed above, Boyd, De Nicolò, and Rodionova (2019) show that they can predict the events method start dates using measures of large declines in bank loans (they call this the systemic bank shock, SBS).\(^{22}\) Using their dating of the starts of crises, Boyd et al. reestimate the probability of a crisis, using the same right-hand-side variables as the other authors. They find that “the probability of a systemic bank shock is neither predicted by the presence of a deposit insurance

---

20. “Mutual liability” means that each surviving bank was liable to pay should another bank become insolvent. “Success” in this context means that the insurance fund was solvent by the start of the Civil War.

21. Panics were stopped despite all the charges of moral hazard that had started in the early 1900s with the state deposit systems. See Robb (1921, pp. 188–89).

22. In the newest version of the paper, dated 2016, the authors look at a number of variables as candidates for measuring systemic bank shock. Earlier versions focused on bank loans.
system ... nor by the generosity of the safety net” (p. 14; italics in original). In other words, they get the opposite result, suggesting “that policy responses to systemic bank shocks are more likely if a deposit insurance system is in place, since the government commitment to honor the associated guarantees is stronger” (p. 12; italics in original).

3.3 Private Regulation

Banks recognized the vulnerability of their debt to runs. And, in the pre-Fed period they formed private bank clearinghouses to clear checks. But, in the clearing process each member bank potentially faced large counterparty exposure to other member banks. This exposure was the origin of their private regulation. But, that was not all. In a crisis, the clearinghouse issued its own money (clearinghouse loan certificates), which were claims on the clearinghouse, the collective membership. This was a kind of insurance as it made the risk of each individual bank irrelevant, and focused on the solvency of the clearinghouse (see Gorton 1985, Gorton and Mullineaux 1987, Gorton and Huang 2003, 2006, and Gorton and Tallman 2018). The prospect of such joint liability was a further incentive to regulate.

Private regulation took many forms, including rules about what information to release or not to release, and so on. For reasons discussed above, most of this was done in secret. Bolles (1903, p. 379): “The extent of the supervision exercised by this [clearinghouse] association over its members the public will never know, because it is best that much of it remain secret. The banks thus associated learn more about one another than they ever would if acting entirely alone, and examinations are made, and warnings given, of which the public has no knowledge. The direct interest that every bank has in knowing the true condition of every other member is one of the great merits of the system.”

Banks were also examined by state and federal examiners, but Smith (1908) described the government bank examinations during this period as “defective” relative to clearinghouse examinations.

The examiners appointed by the Government and the various states, while carrying on their work as best they can, have to labor under difficulties which can be eliminated by this new method (i.e., special examiners appointed by clearing houses), and they cannot make as thorough an examination as the clearing house examiners…. Excessive loans to a firm that is borrowing from several banks, both national and state, can be discovered much better by the clearing house examiner, because he knows the relations of a borrower with the other banks. Each bank might be loaning to this firm what would be safe if he borrowed nowhere else, and in case of trouble all would lose. (pp. 177–78)

In addition, as described by Gorton and Tallman (2018), large banks in the clearinghouse sometimes “bailed out” another large clearinghouse member in a crisis.

Such private regulation, including a form of deposit insurance and bailouts, strongly suggests that the vulnerability of private money to runs is inherent rather than being caused by the government.

For deposit insurance and collateral requirements to work in reducing the likelihood of crisis, regulators must be able to distinguish banks from nonbanks. The
allowable and disallowed activities of banks must be specified. This is done by requiring a charter or license to be a bank that is to issue short-term debt. Nonbanks must not engage in these banking activities. Also, since banks are by design opaque, bank examination has also played an important role.

3.4 Charter Requirements

Because short-term debt is special, entry into banking has (almost) always been limited (less so under free banking). Entry into banking is restricted by the requirement that banks obtain a charter (a license to engage in the “business of banking”), and then banks must satisfy a multitude of requirements, capital and reserve requirements, and also restrictions on activities. Charters can be denied. The activities of banks have been restricted; they are subject to examination. Nonchartered institutions have been restrained from entry into banking.

In the United States, starting with Massachusetts, New Hampshire, and New York, laws were passed in 1799 and 1804 dictating that only banks could issue bank notes. Charters were granted to banks. Initially this may have been due to creating and protecting monopolies.

The New York State Assembly passed “The Act to Authorize the Business of Banking” on April 18, 1838 (the “Free Banking Act”). By this time, the phrase “the business of banking” had a well-established meaning. It referred to “the issuing of circulating notes, receiving deposits, and loaning money on real or personal security” (Curtis v. Leavitt 15 N.Y. 9, 52 (1857)). The business evolved, for example after 1863, when state banks no longer issued circulating notes, but it was still commonly understood what “banking” was. The National Bank Acts were passed after the Civil War and states that corporations formed under it are for the purpose of “carrying on the business of banking” (12 U.S. C. 21). The “business of banking” was transformed over the next century and a half, with for example the introduction of money market mutual funds. The legal and regulatory infrastructure, developed since the 18th century, has not kept up of late.

The phrase “the business of banking” has a long history. In 1797, the British Parliament passed an act to allow the Bank of Scotland, the Royal Bank of Scotland, the British linen Company, and all other banks and banking companies of Scotland which “had exercised the Business of Banking” to continue issuing certain small notes (see Statutes 1798). And earlier, Arnot (1779, p. 532) refers to the “business of
banking.” The Religious Tract Society (1853) informs us that: “The business of banking in the modern acceptance of the term, did not begin in England until the middle of the seventeenth century. It originated with the goldsmiths of London, who, in addition to their trade in bullion, began then both to borrow and lend money. In the former case they gave receipts for the money entrusted to their care, which circulate at their specified value, and thus laid the foundation for the introduction of bank notes” (pp. 114–115). See Gilbart (1866).

The activities of banks were also circumscribed. Typically, banks are not allowed to own equities or real estate. In modern times, banks are generally not able to offer insurance products (though this varies by country). They cannot generally participate in real estate activities beyond acting as a lender.

Limiting entry only makes sense if other firms are not allowed to act like banks by issuing short-term debt. Other firms were “restrained” from entering the “business of banking.” A restraining act is legislation that prohibits nonbanks from issuing short-term debt. For example, in New York the first restraining act, passed in 1804 (“an Act to Restrain Unincorporated Banking Associations”) was passed with the objective of guaranteeing “to banks a monopoly of the rights and privileges granted them, which had been encroached upon or infringed by private associations” (see, e.g., N.Y. Fireman’s Insurance Co. v. Ely, 1825 (2 Cowen, 711)). And again in 1829 New York State enacted Revised (restraining) Statute of 1830 against unauthorized banking, 1 N.Y. R. S. 712, § 6, that says that “no person, association, or body corporate, except such bodies corporate as are expressly authorized by law, shall keep any office for the purpose of receiving deposits or discounting notes or bills, or issuing any evidences of debt, to be loaned or put in circulation as money, nor shall they issue any bills or promissory notes or other evidence of debt as private bankers, for the purpose of loaning them or putting them in circulation as money, unless thereto specially authorized by law.” Once these laws were passed, entering banking depended on having a charter (see Cleaveland 1857 and Symons 1983).

3.5 Bank Examinations

Because banks are by design opaque it is important for private bank clearing houses and later the government to examine bank asset portfolios. Such production of information has always accompanied the issuance of bank charters (and then the examination information is kept confidential). Hammond (1852), for example, states that in New York:

… the law provides for the appointment of bank commissioners, authorized and required frequently to examine into the condition of every bank in the state, and by the aid of the chancellor to suspend the operations of all such banks as they may have reason to believe are making improvident issues (p. 300).

And Willis (1894): “The … law of 1829 directed the appointment of three Bank Commissioners, whose duty it was to visit the banks of the State, examine their condition at least once in every four months, and report annually to the legislature the result of their investigations” (p. 99).
Regulated banks receive subsidies to the extent that entry is limited, deposit insurance is underpriced and there may be ceilings on interest rates paid on bank liabilities. The entry restrictions create monopoly rents for banks, a bank’s “charter value.” Importantly, because charter value is an intangible asset that is lost if the bank fails, it creates an incentive for bank owners to avoid risk which would jeopardize their charter (see Marcus 1990). In this way, charter value helps align the banks’ private propensities for risk taking with the social goal of producing safe short-term debt.

Charter value derives from monopoly rents. There is a literature on bank local deposit monopolies, for example, Diebold and Sharpe (1990), Hannan and Berger (1989, 1991), and Neumark and Sharpe (1992). Bank deposit rates increase far less than one-for-one with the Fed funds rate and vice versa. Drechsler, Savov, and Schnabl (2017) are the latest to document this. They argue that this is the source of monetary policy potency.

The U.S. deposit insurance system worked well over most of its history. But major problems began to appear for regulated banks in the early 1980s due to increases in competition from money market funds (MMFs) and junk bonds. Keeley (1990) argues that these increases in competition resulted in bank charter values declining. This decline led banks to increase default risk through increases in asset risk and reductions in capital. Keeley argues that this drop in franchise value led to increased risk taking in the 1980s, a decade in which the average failure rate for U.S. banks reached a 50-year high of almost 100 per year. Gorton and Rosen (1995) argue that this increase in risk taking was due to managerial entrenchment not moral hazard.

Given that banks always have an option to exit the industry either by moving to shadow banking or leaving banking altogether, charter value plays an especially important role. It provides a benefit, monopoly rents, to entice banks to stay in the banking industry and abide by banking regulations.

4. FINANCIAL SYSTEM METAMORPHOSIS

Granting charters requires knowing what firms are “banks,” firms producing short-term bank debt. Chartering “banks” is a precondition for the success of any bank regulation. Finding the “banks” may seem easy, but historically it has not been so easy. And this will remain a challenge because the financial system—like market economies generally—is always in flux.

The problem is that the financial system changes over time (like other industries) and with it the form of short-term debt can change. In the United States, for example, the private bank note system was followed by the system of national bank notes.

24. Informally, large banks are members of a “club” with the central bank and the club restricts entry, for example, the Canadian banking system. The club is the source of the entry restrictions. And, the case of the United States, for most of its history there were restrictions prohibiting banks from branching across state lines.

25. Saunders and Wilson (2001) argue that charter value is very sensitive to aggregate economic activity. During contractions, charter value can rapidly decline.

National bank notes had to be backed by Treasury collateral, but because Treasuries also have a convenience yield banks did not want to use Treasuries exclusively to back bank notes, so there was an “underissuance” of national bank notes (see Gorton, Laarits, and Muir 2020). The result of this was that demand deposits grew much faster in the United States than in other advanced countries. Demand deposits were the shadow banking instruments of their day. And this was not understood. Economists and regulators could not determine if demand deposits were “money.” Surveys were conducted of the use of demand deposits (see the discussion in Gorton 2012). Surprising as it may seem today, people were confused, despite the fact that the five major banking crises of the U.S. National Banking Era were runs on demand deposits. Hammond (1957), in his Pulitzer Prize-winning book Banks and Politics in America, wrote: “… the importance of deposits was not realized by most American economists … till after 1900” (p. 80).

But, it is not only demand deposits. Until the failure of Lehman Brothers caused runs on MMFs, it was never clear that MMFs were banks. The shares that MMFs sell to the public are legally securities (equity interests in the fund) and so, it was originally determined, MMFs are not banks.

But the net asset value (NAV) of MMFs is designed to be $1 per share. And this is accomplished by only revealing the NAV with a delay (until recent regulatory changes). It is easy for holders of MMFs to redeem their shares dollar-for-dollar by, for example, writing “checks” on their accounts. Are MMFs banks? That was never the question. The Securities and Exchange Commission declared MMF shares to be securities, so the only question was whether banks could offer similar products. The answer was no. In Investment Company Institute v. Camps (401 U.S. 617 (1971)), the Supreme Court held that the Glass–Steagall Act prohibits commercial banks from offering open-end investment accounts, which are virtually equivalent to money market mutual funds. The question of whether MMFs were banks simply never came up.

The final example is “shadow banking.” This term refers to short-term debt issued by wholesale lenders which is backed by debt.28 Examples include sale and repurchase agreements (repo) and asset-backed commercial paper. Unlike with demand deposits and MMFs, there was never any discussion of whether these forms of short-term debt constituted banking. There was never any discussion because the shadow banking system was not actually observed by the regulators and academics. Figure 1 shows the components of privately produced safe debt as a fraction of total privately produced safe debt, from Gorton, Lewellen, and Metrick (2010). Several important points are worth noting. First, the relative decline in demand deposits starts in the late 1970s and early 1980s. Second, the next components are money market instruments (repo, financial commercial paper, and MMFs), followed by AAA

27. “Convenience yield refers to the nonpecuniary component of the return. So, the pecuniary (cash) return can be lower than it otherwise would be. Cash is the leading example. It has no pecuniary return, but it is still held because it is often “convenient.”

28. “Wholesale lenders” refers to the old investment banks and, more generally, to any financial intermediary that finances itself in part by issuing uninsured short-term debt such as repo, commercial paper, of large uninsured certificates of deposit.
securitizations. These two components are shadow banking: short-term debt backed by AAA securitizations.

The figure makes clear that shadow banking is larger than traditional banking. It is also clear that shadow banking is not a recent phenomenon. By 2007, the first year of the financial crisis, this system was at least 25 years old. The banking system had significantly transformed under the noses of regulators and academics. Why did they not see it? The likely answer, provided by Albert Einstein, is that “theory decides what can be observed.”

5. FINAL THOUGHTS

The reality and tragedy is that we have not been able to deliberately design successful bank regulation that delivers long periods of quiet; instead, our regulatory successes have been spillover effects. In the United States, the National Bank Acts which created national bank notes that traded at par were passed to finance the north in the Civil War. Banking panics were not eliminated, but occurred with demand deposits rather than banknotes. Deposit insurance was a populist mandate that passed over the objections of bankers, economists, and initially President Roosevelt, during the Great Depression. In England, the Bank of England gradually learned how to be a central bank, although its origin had to do with government fiscal policy. In Canada, the banking system transformed itself over time to restrict entry; the resulting unintentional creation of charter value mitigated risk-taking incentives.

Why cannot we actively adopt rational bank regulation? First, there is an intellectual failure. Prior to 2007–8 economists, especially macroeconomists, dismissed
financial intermediation as a relevant topic for study. Financial crises were not seen as inherent to market economies and it was unconsciously assumed that a highly developed economy had overcome this problem. So, we simply have not learned enough.

Second, without seeing the structural problem with short-term debt, the focus has been on the government as the cause of crises: moral hazard and too-big-to-fail. But, these words do not explain why there was no crisis in the United States between 1933 and 2007. Nor have their proponents explained why crises now are different from those in the past when there was no central bank and no deposit insurance.

Third, the political economy of bank regulation is such that new regulation is passed quickly after a financial crisis. But, historically, as well as in the example of 2007–8, the immediate postcrisis environment is fraught with antibanker tension. The idea that there is a flip-side to the invisible hand, where things can go wrong and market participants must bear some costs, is not a view that carries much weight in the public forum. Postcrisis periods are littered with failed bank reform.

Fourth, a basic issue is that banks are private companies. They cannot be forced to do things they do not want to do. They can exit. Capital can always migrate, albeit over a longer time horizon than most regulators and academics think about. Money market mutual funds, for example (a form of regulatory arbitrage), developed over 30 years or so. In the end, regulations only determine where these private companies operate. This is why charter value is important, to provide incentives for capital to remain in banking.

LITERATURE CITED


The Statutes at Large, from the thirty-fifth Year of the Reign of King George the Third to the Thirty-eighth Year of the Reign of King George the Third, inclusive (1798) (George Eyre and Andrew Strahan, Printers to the King’s Most Excellent Majesty).


