THE PARADOXICAL GENESIS OF TOO-BIG-TO-FAIL

Thomas S. Umlauft*

Abstract

At least since the Global Financial Crisis of 2007-2009, the problem of too-big-to-fail (TBTF) has received widespread attention. The research conducted in this context has, however, generally focused on the econometric aspect and the contribution of the TBTF doctrine to the financial crisis of 2007-2009, while the economic historical approach has been confined to tracing the doctrine to its first appearance. This paper attempts to fill this gap in the academic literature by offering an explanation for why, as opposed to how, the TBTF doctrine has developed. This paper identifies the US population’s distrust and at times hostility against the prospect of concentration of power in large financial institutions as the causal factor leading to the TBTF phenomenon. The resulting socially non-optimal regulation favoured a fragmented and fragile banking system based on small unit banks at the cost of more diversified branch banks. The Great Depression impressively highlighted the deep structural flaws of the US banking system. At the same time, however, it caused a shift in the public opinion, which had generally been opposed to deposit insurance, and thereby aligned the public interest with that of small banks, which would profit most from deposit insurance. The newly acquired public and political support enabled weak unit banks to lobby successfully against reforming the banking structure and instead for the adaption of federal deposit insurance. However, the Federal Deposit Insurance Corporation (FDIC) only addressed the symptoms of the weak banking industry but not its causes. Moreover, the strongly biased FDIC policies have generally favoured creditors at large banks, which ultimately led to the TBTF doctrine which, in turn, provided banks with a non-technical incentive to grow in size in order to gain TBTF protection. Initially aimed at preserving the US financial landscape based on small unit banks, the FDIC as the main conduit for TBTF rescues thus became the main driver for big bank corporate welfare. Deposit insurance gave rise to TBTF and, at the same time, put small banks deemed “too-small-to-safe” at a competitive disadvantage, further accelerating the trend towards increasingly large and complex banks.

JEL classification: D72, G01, G18, G21, G28, G33, I51, N11, N21, N22

Keywords: Deposit Insurance, Early Deposit Insurance Systems, Too-Big-to-Fail, Banks, Bank Failures, Economic Theory of Regulation, Capture Theory, Rent-Seeking, Regulation, Lobbying, Moral Hazard

* Department for Economic and Social History, University of Vienna
E-mail: a0305256@unet.univie.ac.at

1 Introduction

The Global Financial Crisis of 2007-2009 is closely connected to the failure of large banks which, owing to their size, were deemed to be too-big-to-fail (TBTF). Although smaller banks failed in large numbers as well, the causality is generally seen as running from large to smaller banks (Bair 2000; Boyd & Heitz 2012, p. 2; Slovik 2012, p. 9). Therefore, the interest in the too-big-to-fail problem in the wake of the Global Financial Crisis has grown dramatically and attracted a large number of studies. The studies undertaken have, however, mainly focused either on the econometric aspects, i.e. quantifying TBTF subsidies (Baker & McArthur 2009; Haldane 2010; Noss & Sowerbutts 2012; Ueda & Weder di Mauro 2012), or the contribution of the TBTF doctrine to the Global Financial Crisis (see, for example, Boyd, Jagannathan & Kwak 2009, Umlauft 2014). Historical approaches have usually concentrated on the development of the TBTF doctrine per se, that is the development from its first appearance, but not its underlying causes (see, for example, Shull 2010). This paper therefore attempts to fill this gap in the academic literature by addressing the question why, as opposed to how, TBTF has developed. This paper in hand shows that the underlying factors that have led to the TBTF doctrine are rooted in the historical development of the United States, whose population as well as politicians have traditionally exhibited a marked distrust and at times hostility against the prospective of concentrated power in large financial institutions. This distrust, combined with the opportunities it offered for banking institutions to
lobby for favourable legislation, significantly shaped banking sector regulation. As a result, the regulation of the financial system allowed for a fragmented banking system, predominantly based on small, single-unit banks, prone to financial crises. The Great Depression impressively highlighted the deep structural flaws of the US banking system. Yet, the response to the large number of bank runs and banking panics during 1929-1933 was not so much a reform of the dysfunctional system but rather addressing its symptoms (bank runs, banking panics, bank failures) by introducing federal deposit insurance in 1933. Insurance of deposits by the government, however, aggravated the structural problems by subsidising banks’ risk-taking. Therefore, the establishment of the Federal Deposit Insurance Corporation (FDIC) not only did not address the underlying structural problems of the US banking system but also increased its fragility by subsidising risky behaviour. Beginning in the 1970s, the FDIC’s approach to the resolution of the increasing number of bank failures became successively more biased – while small banks generally were liquidated protecting depositors only within the insurance limit, large bank resolutions usually protected creditors and depositors in full. Insofar as the FDIC encouraged risky behaviour of banks and constitutes the main conduit through which TBTF measures have been undertaken, federal deposit insurance both precipitated and made possible TBTF.

The remainder of the paper is structured as follows: Part 2 will discuss the methodological framework used in order to answer the research question and give a short review of the related theoretical literature. Finally, the applied techniques as well as the sources and data used will be discussed. Part 3 will present a narrative analysis of the problem by briefly reviewing the evolution of the US financial system with regard to important developments connected with the research question. Part 4 will provide a conclusion, summarising the main findings in this paper and their implications and offering policy recommendations directly derived from these.

2 Methodology & Review of Theoretical Literature

The paper draws heavily on the theory of regulation. Until the middle of the 20th century, regulation had primarily been explained by the public interest theory, which holds that regulation results from the demand of the public and is aimed at correcting failures and inequities of unregulated, laissez-faire markets. Therefore, governments were seen to be both capable of and willing to efficiently eliminate market failures and thereby raise social welfare. As Posner (1974) notes, the public interest theory had mostly been implicitly assumed rather than articulated. The economic theory of regulation, first advocated by Stigler (1971) and Posner (1974), provides an economic rationale for regulation by arguing “that regulation is supplied in response to the demands of interest groups struggling among themselves to maximize the incomes of their members.” (Posner 1974, p. 335-336) The economic theory of regulation is based on Olson’s (1965) theory of “Collective Action” which posits that the smaller the group and the higher its per capita stake at the outcome, the more likely it will succeed in influencing regulation. Accordingly, the supply of regulation is administered by the same factor as other goods: demand. Since industries are directly affected by (costly) regulation they have a larger stake at favourable outcomes than the widely dispersed public. Because industry groups possess greater resources, they often succeed in instrumentalising the state’s coercive power in order to assure favourable regulation at the cost of the public or less influential industries. Insofar as interest groups often succeed in influencing regulators to enact favourable regulation to maximise their profits by acquiring economic rents, the economic theory of regulation is often referred to as capture theory. Laffont & Tirole (1991) supplement Stigler and Posner’s framework by enhancing the theory from a theory focused on the demand side by providing a supply side aspect. Laffont & Tirole thereby significantly increase the theory’s credence as well as its explanatory power. The authors allow for agency problems15 that manifest themselves in favourable regulation owing to benefits for regulators by regulated firms. According to Laffont & Tirole, decision-makers may be captured by interest groups with monetary bribes, revolving doors between regulating agencies and the industry and personal relationships as well as political contributions.

A different strand of academic literature critiques the economic theory of regulation for focusing on self-interest alone and rejects it as too narrow. Specifically, it is argued that ideology and altruism may play an important role in legislation. On a purely logical level, Kau & Rubin (1979) present a convincing case by pointing out that economists, more than any other group of social scientists, are more likely to reject government intervention in favour of free markets. Yet, the authors continue, if such policies were to be adopted, the income of economists would fall, since the government constitutes a significant demand factor for economists. Thus, Kau & Rubin conclude,

14 Sprague (2000) gives a detailed account of the early history of too-big-to-fail bailouts, highlighting the leading part of the Federal Deposit Insurance Corporation in TBTF rescues.

15 Principle-agent relationships arise when principals (e.g. owners, electorate) delegate powers to agents (e.g. managers, politicians) in order for the latter to act in the interest of the former. Agency problems are the result of discrepant interests of the two parties and denominate actions which benefit agents at the cost of principals.
economists advocate a policy on ideological grounds, detrimental to their self-interest. Moreover, the authors argue that economists seem to implicitly accept the notion that ideology does matter with regard to legislation as most articles in the field of economics offer some sort of policy advice. Clearly, the raison d’etre of advice is predicated on the influence of ideas. In subsequent research, ideology is found to be a dominant factor in explaining voting behaviour related to minimum wages in the United States (Kau & Rubin 1978) and on coal mining laws (Kalt & Zupan 1984). Goldstein (1988) finds similar evidence related to protectionist US trade policies where redistributive motives seem to have played an important role, and Berglöf & Rosenthal (2000, 2003) suggest that ideology is a key determinant for explaining voting patterns on US bankruptcy legislation between 1800 and 1976. Ideology is also found to be a dominant factor in explaining regulation in the financial sector. Roe (1994), for example, argues that the population’s fear of concentration of power within financial firms substantially shaped banking regulation in the United States. Similar interpretations can be found in Viner (1936) who argues that the fear of Wall Street and financial conglomerates shaped policy-making in a variety of ways.

Finally, somewhat related to both the economic theory of regulation and regulation based on ideology is what Buiter (2008, p. 106) calls “cognitive regulatory capture”, which is

“not achieved by special interest buying, black-mailing or bribing their way towards control of the legislature, the executive [...] or some other important regulator or agency [...] but instead through those in charge of the relevant state entity internalising, as if by osmosis, the objectives, interests and perception of reality of the vested interest they are meant to regulate and supervise in the public interest.”

However, the explicit models in the above cited articles, with the notable exception of Buiter (2008), who uses the concept in a qualitative way, all suffer from the common weakness of monocular explanations. Given that measuring ideology is a complicated if not impossible task, models that link voting behaviour to some form of measurement of ideology are of limited use only. The same, of course, is true for economic incentives whose measurement is equally problematic, given that many of the factors put forth by Laffont & Tirole (1991) in order to explain voting behaviour – bribes, future employment prospects, personal relationships, etc. – are strictly not observable due to them being either illegal or potentially damaging for politicians’ reputation. Moreover, there probably is some overlap of the vested interests’ of lobbying groups and captured politicians due to humans’ tendency to morally justify one’s behaviour. The task of measuring either function is further complicated by the observed phenomena of shirking\(^\text{17}\) (see Rowley & Schneider 2004, p. 293-294 for an overview) and log-rolling\(^\text{18}\) (Irwin & Kroszner 1999). Finally, Poole & Rosenthal (1994) stress the difficulties resulting from party discipline-influenced voting behaviour. These problems render monocular explanations not a particularly promising approach for a positive analysis of history. Although the above mentioned models provide the conceptual framework for the analysis conducted in this paper, a narrative approach will be applied in order to fully account for the various factors at work – ideology, private interests as well as cognitive regulatory capture – and in particular their intertwined occurrences. The paper therefore draws on econometric studies, narrative analyses and official releases as well as contemporary accounts.

In view of the financial crisis of 2007-2009 and subsequent efforts to reform the financial sector (e.g. Dodd-Frank Act, Basel III), historical bank regulation in the United States offers a particularly promising field for demonstrating how ideology and private interests may interact and lead to inefficient outcomes that benefit a small group at the expense of more widely dispersed groups and ultimately the economy.

3 Descriptive Analysis

Banks play a vital role in any economy due to their primary purpose of channelling funds from surplus units to those with deficits. In other words, banks and other financial intermediaries make sure that the resource money does not lie idle but instead is directed toward a constructive use. As distress in the banking sector is generally associated with severe output losses (Friedman & Schwartz 1963, Bernanke 1983; Boyd, Kwak & Smith 2005), banks and other financial intermediaries are heavily regulated in most countries (Allen & Gale 2001). Compared to other countries, the US financial system is special on many accounts. The two most important characteristics that historically (hence) defined the US banking sector are (1) the dual banking system\(^\text{19}\) and

\(^{16}\) However, a case could be made in favour of rationally acting economists inferring as policy advice may be aimed at incentivising disadvantaged groups so that they organise and thereby provide a counter-measure for influential, concentrated interest groups.

\(^{17}\) Shirking constitutes a principal-agent problem and arises when politicians’ voting behaviour does not represent the ideology of their electorate but instead the politicians’ own views.

\(^{18}\) Log-rolling occurs when groups with unconnected interests form supportive voting coalitions in order to obtain favoured legislations.

\(^{19}\) The dual banking system describes the United States’ bank chartering both under state and federal law. While banks chartered under state law (state banks) are supervised by the respective state legislature, federally chartered banks (national banks) are supervised and regulated by the Office of the Comptroller of the Currency (OCC).
(2) unit banking\textsuperscript{20}. Both characteristics are unique to the United States\textsuperscript{21} and will be outlined in the following two chapters.

\textbf{Dual Banking}

The dual banking system is rooted in the United States Constitutional Convention in 1787, when federalists and anti-federalists\textsuperscript{22} debated over whether the power to regulate the financial sector should belong to the states or the federal government. In 1791 the federalists seemed to have won the day by establishing the federally chartered \textit{First Bank of the United States} (BUS1), which acted as a quasi-central bank by systematically converting banknotes from state banks and thereby preventing over-issuing.\textsuperscript{23} In addition to that, its chartering was also intended to ensure government funding. However, the BUS1 charter was not renewed in 1811 which left a void that was filled by small state-chartered banks which rapidly increased after the dissolution of the BUS1. From 1811 to 1816, the number of state banks increased from 90 to 260, while the outstanding value of banknotes rose from $28 million to $68 million. The resulting over-banking after 1811 combined with the war against England forced wide areas of the US banking system to suspend convertibility of banknotes into species. The disastrous banking situation convinced many former opponents of federally chartered banks to charter the \textit{Second Bank of the United States} (BUS2) in 1816 (Myers 1970, p. 78-84). Jenkins & Weidenmier (1999) show that voting behaviour on BUS2 is explained by ideological as well as economic factors. On the one hand, beliefs about how much power the federal government should have in relation to state governments constituted a divide. This ideological motive in turn was often shaped by economic interest of the states as state bank revenues provided a significant portion of state revenues (Sylla, Legler & Wallis 1978; Jenkins & Weidenmier 1999).

Again, the disciplining presence of the BUS2 prevented state banks from imprudent note issuance, and again, after the charter expired in 1836, it was not extended, this time due to a veto by President Jackson, a well-known opponent of banks in general.\textsuperscript{24} With the power to regulate banks back solely in the hands of the states, 1837 marks the beginning of the so-called free banking era, which by many is seen as a response to the concentration of power. In free banking systems, states lost their discretionary chartering privilege in favour of a specified set of rules whose satisfaction automatically granted a bank charter. Some researchers attribute the move towards free banking to a democratisation of banking which was aimed at stripping governments of their often misused power of chartering towards a more rule-based and objective approach (Bodenhorn 2006). Others see free banking merely as a consequence of decreased charter values which made discretionary chartering by governments not worthwhile anymore (Grossman 2010, p. 230).

The Civil War (1861-1865) provided the federal government with the opportunity to reform the banking and financial system because the southern states, which had traditionally been opposing federal regulation, had withdrawn from Congress (Komai & Richardson 2011, p. 3).\textsuperscript{25} A crucial factor for understanding the reformation of the banking sector was the federal government’s need for funding in order to finance its war efforts. The National Banking Act of 1863, which was modified the following year by the National Banking Act of 1864,\textsuperscript{26} finally brought the federal government into banking by establishing federally chartered national banks as an alternative to state banks. Except for the short periods between 1791 and 1811 and between 1816 and 1836, when the \textit{First} and the \textit{Second Bank of the United States} operated under a federal charter, this was the first time that banks from different states would operate under a uniform charter, regulation and supervision. In 1862, Congress had granted national banks a quasi-monopoly on the issue of banknotes by levying a 2% tax on banknotes issued by state banks. To further encourage bank chartering under federal law, the Revenue Act of 1965 increased the tax to 10%, which ultimately rendered the issuance of notes by state banks unprofitable (Myers 1970, p. 163; Committee on Branch, Group, and Chain Banking 1932, p. 61). As banknotes issued by national banks were required to be backed by US government bonds, a much needed demand for government debt was

\textsuperscript{20} Unit banks are comprised of only one office and no branches.

\textsuperscript{21} See Federal Reserve System (1924: 929); Calomiris & White (1994, p. 148).

\textsuperscript{22} The terms federalism and anti-federalism in the United States have repeatedly undergone a change in meaning. Although today federalism generally denotes the strive for strong regional administrations as opposed to the central government, in the context of this paper, federalism will be used in its historical meaning to describe proponents of a strong central (federal) government.

\textsuperscript{23} The Federal Reserve System (Fed), the United States’ central bank, was only established in 1913 under the impression of the Crash of 1907. Over-issuing of bank notes had been a continuous problem until 1863 when a uniform currency was introduced. Until then, note issuing powers had belonged to banks.

\textsuperscript{24} For a historical account of both the \textit{First} and the \textit{Second Bank of the United States}, see Holdsworth & Dewey (1910) and Myers (1970, p. 66-94).

\textsuperscript{25} In fact, opposition to federal government, mainly from the western and southern states, had prevented the United States from establishing a central bank, a uniform currency and nationwide bank regulation. Besides establishing national banks, the National Banking Act also introduced a uniform currency backed by the United States’ Treasury (Komai & Richardson (2011, p. 3-4).

\textsuperscript{26} The National Banking Acts of 1863 and 1864 initially were called National Currency Acts of 1863 and 1864 because besides creating the national banking system the legislation introduced a uniform national currency backed by US bonds and the US government.
created that would help finance the war efforts of the government (Myers 1970, p. 163). However, state banks did not disappear and in fact flourished from the 1880s onwards. Spong (2000, p. 19) attributes this trend to two factors: First, the importance of banknotes decreased significantly vis-à-vis deposits in the second half of the 19th century, stripping national banks from one of their main advantages over state banks (Around the middle of the 19th century the ratio of deposits to banknotes had been about 1:1. However, by 1870, the ratio had increased to 2:1 and by the end of the century to 7:1 (FDIC 1998: 12; Golembe & Warburton 1958: I-2)). Second, the yields on government debt eligible for note backing began to fall in the 1880s and continued to do so until the end of the century (A time-series analysis of Homer & Sylla (2005, p. 283-285) of US government yields indicates that yields on US government debt had been decreasing since the 1860s). Another factor for the continuing importance of state banks may well have been the prohibition of branching for national banks. Since certain states allowed branching for state banks, this constituted a major advantage for state banks as the Banking Acts of 1863 and 1864 prescribed unit banking for national banks (Myers 1970, p. 165) (The National Banking Act of 1864 dropped the plural of “office” as used in the Act of 1863 (Myers 1970, p. 165). Insofar as “branching” had not been used in the Act of 1864 (Committee on Branch, Group, and Chain Banking 1932, p. 51), the revision most likely aimed at eliminating an ambiguous passage).

Summarising, with the notable exception of the First and the Second Bank of the United States, until 1863 bank chartering, regulation and supervision was left in the hands of the states due to a struggle between federalists and anti-federalists. Only in 1863, facilitated by the Civil War, the federal government managed to assume a significant role in the banking industry. However, the newly created national banks did not succeed in supplanting state banks, leading to the co-existence of state and national banks henceforth (see figure 1 for the evolution of the number of state and national banks) and therefore to the dual banking system characteristic for the United States’ financial system.

**Figure 1.** Number of National and State Banks, 1863-1928. Data: Comptroller of the Currency (1932, p. 3).

### 3.1 Unit Banking

A direct consequence of the United States’ dual banking system is its equally unique system of unit banking. Until recently, banks with only one office and no branches dominated the United States’ banking landscape.²⁷ Calomiris & White (1994, p. 148) argue that one of the most important preconditions for bank fragmentation was the Supreme Court’s decision not to apply the commerce clause to banks.²⁸ As a corollary, state

---

²⁷ Interstate branch banking was only introduced in 1994 when the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994 repealed the McFadden provision on interstate banking.

²⁸ The commerce clause (Article 1, Section 8, Clause 3 of the U.S. Constitution) grants Congress the power “to
governments’ authority to charter banks allowed individual states to design their own banking systems which systematically protected domestic banks by limiting competition. While this approach explains the absence of interstate banking, it does not explain the prohibition of intrastate branch banking. However, White (1984, p. 1086-1089) links unit banking to populist propaganda of single office bankers to capture rents by effectively arousing fears of branch banking, and Calomiris (2000) regards farmers as beneficiaries of unit banking inssofar as branching restrictions tied banks to specific locations and therefore secured loans for the agrarian population. Moreover, Bodenhorn (2003, p. 12-18) demonstrates that bank chartering by states often involved bribes and favours like government loans in exchange for charters. It is therefore conceivable that chartering a higher number of unit banks was perceived to be more profitable by politicians than chartering a significantly lower number of branch banks.

The main advantages of branch banking have already been pointed out by Sprague (1903, p. 243): (1) superior ability of management, (2) decreased risk through the enhanced possibility of diversification of investments and (3) greater effectiveness of reserves. While the first point is based on the logic that larger firms are able to attract, both by their reputation and the possibility of paying higher salaries, more talented managers, the two other points are firmly rooted in both basic financial theory and mathematical proof. Consequently, Sprague (1903, p. 242) asserts that “[u]pon few subjects has the consensus of opinion of both economists and financial writers been more general than upon the advantages of branch banking over a systems of separate local banks.”

Indeed, a wide range of scholars see unit banking as an important factor for banking system instability in the United States. Calomiris (2000) shows that systems with large, geographically diversified banks were less prone to panics, which had a lower failure probability and incurred smaller losses when they failed. Furthermore, systems with branch banking, while not immune to panics, recovered more quickly than unit banking systems ceteris paribus. Bordo (1985, p. 27) links the nationwide banking panics between 1870 and 1933 to the United States’ unit banking system “in a period when they were an historical curiosity in other countries.” Similar opinions can be found in White (1984) who compares the US’ to the Canadian banking system which proved to be much more resilient to panics than the US system.

The fragility of unit banking has long been recognised and a variety of measures were undertaken in order to stabilise the very same. Most notably, bank obligation insurance schemes have been adopted in two waves by individual states in order to protect the payment system from the disruptive effects of bank failures (Golembe 1960, p. 195; Calomiris & White, p. 148). The first bank obligation insurance scheme was introduced in 1829 in New York. Under the New York “safety fund”, banks were required to pay an assessment into an insurance fund which would reimburse creditors of failed participating banks (Bodenhorn 1996). Vermont and Michigan soon adopted similar bank insurance schemes in 1831 and 1836, respectively. Concurrent with this, Indiana introduced a bank insurance system in 1834, which made member banks mutually liable for the outstanding debt of other member banks. In order to limit risk-taking, supervision was assumed by bank directors whose interest was to prevent failures of other banks due to the shared liability. Mutual guarantee of bank obligations was also introduced 1842 in Ohio and 1858 in Iowa. The bank obligation guarantee systems in Indiana, Ohio and Iowa all dissipated in 1866 when the prohibitive tax on banknotes issued by state banks forced state banks to convert to national banks, but no creditor had lost any money under these systems (Golembe 1955). The deposit insurance experiments of New York, Vermont and Michigan, on the other hand, all collapsed under the weight of bank failures. Uniform assessment fees not based on risk as well as the removal of market discipline had provided banks with ample opportunities to engage in unsound practices. Bank failures in these states therefore were frequently attributable to risky behaviour or outright fraud of insured banks (Golembe & Warburton 1958, Bodenhorn 1996, Garlock 1926).

The second wave of deposit insurance schemes was triggered by the panic of 1907. Between 1908 and 1929, eight states introduced deposit insurance schemes, all of which were based on assessments paid into a fund (The second wave of bank insurance

29 Bernoulli is generally credited for proving the law of large numbers, while the ground-breaking work of Markowitz (1952) on portfolio selection demonstrated that diversification may decrease risk with no negative effect on return. However, the proverb that warns us from putting all eggs in one baskets suggests that the concept and benefits of diversification has been well-known for quite a while.
schemes only insured deposits because the Banking Act of 1863 had made banknotes secured by US Treasury bonds and convertible at face value into specie at the Treasury, therefore rendering insurance of banknotes unnecessary). As a consequence of implementing insurance systems that lacked an alignment of interests by subsidising risk-taking, all insurance funds failed in the 1920s, leaving creditors largely unprotected. Deposit insurance systems in the early 20th century had been adopted by states where state legislature had firmly established unit banking laws which had promoted a large number of small, undiversified banks. Therefore, deposit insurance was more likely to be adopted in states where small unit banks were dominant and hence where deposit insurance’s assumed effect of stabilising the banking sector was most needed due to the weak banking system. Thus adoption of deposit insurance was seen as a means of addressing the symptoms of weak unit banking systems (bank failures), without having to reform the banking system (Colomiris & White 1994, White 1981). However, while state deposit insurance systems did decrease the number of suspensions due to runs, failures due to mis-management and risk-taking increased (Chung & Richardson 2006). Moreover, Dehejia & Lleras-Muney (2007) show that the expansion of the financial sector resulting from deposit insurance adversely affected economic growth, while increased banking activity induced by branching had positive effects on growth.

Notwithstanding the widely recognised advantages of branch banking and the failure of bank obligation insurance to stabilise unit banking systems, by the end of the 19th century unit banking had been made the norm. This was facilitated by the prohibition of national banks to operate branches as prescribed by the National Banking Acts of 1863 and 1864. In 1895, only four states therefore allowed branch banking (Grossman 2010, p. 238). The early 20th century saw an increasing interest in branch banking, especially after the post-war agricultural depression, which had weakened many rural unit banks (White 1985, p. 1085). In the period from 1900 to 1925, the number of branches rose from 119 to 2,525 (Calomiris & White 1994, p. 151) as the number of states permitting some kind of branching more than doubled from less than ten to more than 20 (Grossman 2010, p. 238). The development of the share of banks operating branches is depicted in figure 2 below, demonstrating the increase of branching activity of US banks, albeit at a continually marginal level (Cf. Committee on Branch, Group, and Chain Banking (1932, p. 6); Comptroller of the Currency (1931, p. 3)).

**Figure 2.** Share of State & National Banks Operating Branches, 1863-1928

![Figure 2](image)

Data: Committee on Branch, Group, and Chain Banking (1932, p. 6); Comptroller of the Currency (1931, p. 3)

The trend towards more liberal branching legislation and the threat it posed to unit banks in the 1920s motivated the latter to actively lobby against branch banking as they pushed the American Bankers Association to declare:

>“We regard branch banking or establishment of additional offices by banks as detrimental to the best interests of the people of the U.S. Branch banking is contrary to public policy, violates the basic principles of our government and concentrates the credits of the
nation and the power of money in the hands of a few. “ (Economides, Hubbard & Palia 1995, p. 23)

As can be seen, the typical stereotypes were used in an attempt to instrumentalise the public and politicians for the vested interest of unit banks. As a corollary, the trend towards more liberal branching in state legislature was counteracted by the McFadden Act in 1927, which prohibited branching and therefore codified the US unit banking system – or at least severely constrained branch banking – by explicitly interdicting interstate branching for federally-chartered banks, although it did allow intrastate banking for national banks to the extent permitted by state legislation within the city limits of parent institutions (Committee on Branch, Group, and Chain Banking 1932, p. 4). Not surprisingly, Economides, Hubbard & Palia (1995) provide statistical evidence that politicians from states with a large proportion of unit banks voted in favour of the McFadden Act and hence branching restrictions, indicating successful lobbying on the part of unit banks, concluding that the Act benefited small unit banks which would not have been able to compete with large branch banks without branching restrictions.

3.2 The Great Depression

The 1920s and the Great Depression highlighted the deep structural flaws of the US banking system. From the beginning of the 1920s to 1933, the total number of banks in the US fell from almost 30,000 to less than 14,000 (Upham & Lamke 1934, p. 247). The seminal work “A Monetary History of the United States” by Friedman & Schwartz (1993 [1963]) stresses that bank failures during the Great Depression resulted from panics, in which depositors withdraw money indiscriminately and therefore regardless of banks’ fundamentals. The authors therefore conclude that the large number of bank failures were due to liquidity, not solvency issues and the failures need not have happened if the Federal Reserve had provided liquidity support for distressed banks. As deposit insurance, similar to the lender of last resort function of central banks, can be seen as a liquidity insurance, even Friedman (1960, p. 21), whose attitude towards government intervention usually is not as favourable, praises “deposit insurance [as] the most important structural change in our monetary system in the direction of greater stability since the post-Civil War tax on state bank notes”.

The notion that a lack of liquidity stood at the core of the banking panics during 1929-1933 was shared by many contemporary politicians and policy-makers. If liquidity were to blame for the Great Depression failures, then, indeed, a strong case could be made for deposit insurance. Federally insured deposits would make deposits less likely to be withdrawn in times of crises and therefore would forestall “contagion” and a domino effect during which solvent but illiquid banks would fail.34

However, more recent research suggests that contagion was not an important factor for the high number of bank failures during the Great Depression. In contrast to Friedman & Schwartz (1993 [1963]), Wicker (2000) finds that the Great Depression panics were mostly regionally confined and hence no nationwide phenomena. Wicker confirms the four banking panics identified by Friedman & Schwartz between 1930 and 1933, although with slight modifications regarding their exact dates, and finds a fifth “mini panic” confined to the city of Chicago. Unlike Friedman & Schwartz, Wicker employs disaggregated data using Federal Reserve District data of Federal Reserve Notes in circulation as a proxy for depositor confidence in the banking system. Complementary, Wicker attempts a micro-history of the areas most affected by the banking panics.35 According to Wicker, the banking panics constituted separation equilibria, not pooling equilibria, in as much as panics were confined to specific regions and bank runs were directed towards specific banks which were known to be weak.

The first panic from November 1930 to January 1931 was concentrated to the South and targeted banks which were known to be in some business relationship with Caldwell and Company in Nashville, Tennessee, the largest investment bank in the South. Wicker provides evidence that three of the four banking panics constituted separation equilibria. That is, panics were confined to specific banks which had problems or were insolvent. The first panic from November 1930 to January 1931 was concentrated to the South and targeted banks which were known to be in some business relationship with Caldwell and Company in Nashville Tennessee, the largest investment bank in the South which failed in November 1930. In contrast to Friedman & Schwartz, who identified the failure of the Bank of the United States as the initial and primary cause of the panic, Wicker emphasises the importance of the demise of Caldwell, which controlled the largest Southern chain of banks with assets above $200m as well as the largest insurance group of the South with assets of $230m. Therefore, “[t]he failure of Caldwell and Company had immediate repercussion in four states, namely Tennessee, Kentucky, Arkansas and North Carolina in the Atlanta, St. Louis, and Richmond Federal Reserve Districts.” (Wicker 2000, p. 33).

Wicker points out that most bank failures during the first banking panic can be attributed to their

---

34 Since long-term loans are not easily convertible into cash without incurring substantial losses to the bank, unwarranted withdrawals may render solvent but illiquid banks insolvent. In other words, perceived insolvency may lead to a self-fulfilling prophecy insofar as an ex ante solvent banks may become insolvent due to forced liquidation.

relationship with Caldwell, indicating a separation equilibrium, not a pooling equilibrium. Wicker concludes:

“The principal conclusion to emerge from our description and analysis of the accelerated bank suspensions in November 1930-January 1931 is that there was no national banking crisis. The banking difficulties were region specific, that is, of some local and regional concern but without national importance.” (Wicker 1996, p. 58)

Although Wicker is unable to identify the initial shock of the second panic from April to August 1931, the subsequent two shocks are clearly attributable to Chicago in June and in Toledo in August. The second panic was thus region-specific as well and does not seem to have been motivated by irrational deposit withdrawals. Instead, the Chicago runs were concentrated on newly-chartered, small banks which had provided funds for real estate developers, again indicating a separation, not a pooling equilibrium.

"From our examination of the second banking crisis we conclude that there was no geographically diffused nationwide banking crisis between April and August 1931. Banking suspensions were centered in a relatively few Federal Reserve Districts. The banking crisis was largely a region specific phenomenon; it had not yet become a full-fledged national crisis." (Wicker 1996, p. 72)

Similarly, bank failures during the third panic from September to October 1931 were confined to Chicago, Pittsburgh and Philadelphia and predominantly affected savings banks and trust companies in the periphery. Wicker (2000, p. 99) notes:

“Our analysis would tend to suggest that the waves of bank suspensions in September – October do not appear to have conformed fully to the conventional view of a banking panic; that is, there was no indiscriminate run on banks by depositors whose confidence in banking institutions in a given area had been shattered. Bank runs, especially among urban banks, appear to have been directed against particular banks that were known to be weak.”

Only the fourth panic from February to March 1933 indicates some degree of pooling behaviour, as both solvent and insolvent, sound and unsound banks became subject to runs. The panic originated in Detroit due to the impeding collapse of the Guardian Group, which had been heavily exposed to real estate, as Michigan declared a bank holiday. According to Wicker (2000, p. 121) “[t]he declaration of the Michigan holiday spread fear and uncertainty quickly to the contiguous states who promptly placed restrictions on deposit withdrawals.” Subsequent bank moratoria in 48 states “brought banking operations to a virtual standstill.” (Wicker 2000, p. 148) Hence, Wicker identifies the attempt to forestall panic and contagion as the primary factor causing the very same, which prompts Wicker (p. 48) to note that “[t]he banking panic of 1933 remains an anomaly among US financial panics.”

Calomiris & Mason (2000) show that fundamentals explain most failures and therefore that contagion or liquidity played a relatively unimportant role in bank failures and banking panics of the Great Depression until 1933. Calomiris & Mason (1997) find that the characteristics of banks which failed in Chicago during the panic of 1932 were not different from those before the panic. Therefore, the failures during the panic of 1932 reflected the weakness of failing banks, while solvent banks generally withstood the panic. They find that contagion did not play a major role in the Chicago banking panic of 1932. Similar findings are presented by White (1984) who demonstrates that US-wide bank failures during the 1930 panic were similar to those of the 1920s and not attributable to banking panics. Postel-Vinay (2013) analyses the banking system in Chicago during the years 1923-1933 and finds that, although banks seem to have suffered from indiscriminate runs, the ones which failed the earliest in the 1930s held a higher share of illiquid, risky assets.

Instead of contagion, many scholars attribute the large number of bank failures during the Great Depression to the prevalent unit banking at that time and argue that, had branch banking been allowed, there would have been fewer failures (Bordo 1985, White 1984, Calomiris 2000). Wheelock (1993) provides empirical evidence that bank failure rates during the 1920s were closely correlated with excess capacity, or overbanking, as measured by banks per capita. The number of banks per capita, in turn, was highest in states that restricted branch banking. Cross-sectional studies on bank failures by and large support the notion that states with branch banking were more stable than unit banking states (see Wheelock 1995; Mitchener 2004 for differences in the US and Grossman (1994) for international evidence). Moreover, also contemporary economists and financial journalists had long been stressing that unit banking was an important factor contributing to the United States extraordinary fragile banking system (Sprague 1903; Viner 1936).

However, small unit banks that were not able to compete with larger urban branch banks, had long been supporting deposit insurance. The Great Depression and the resulting loss of confidence in the

---

(36) Somewhat surprising, studies based on intrastate data find that branch banks were more likely to fail than unit banks. Carlson & Mitchener (2005) resolve this puzzle by suggesting that branch banking led to increased competition. The result, according to the authors, was a stronger banking system (which is supported by interstate comparisons) without branch banks necessarily being the strongest banks (as suggested by intrastate data). Carlson & Mitchener (2009) confirm this finding by examining the Californian banking system as a pioneer of large-scale branch banking during the Great Depression. They find that small banks responded to entries of large branch banks by adjusting their operations and increasing their efficiency.
banking sector by the public provided the former with a strong public support for deposit insurance on a federal level. Calomiris & White (1994) argue that weak unit banks had long been in favour of deposit insurance but had lacked the political stamina to lobby against larger, urban branch banks which traditionally had opposed deposit insurance. The Great Depression, however, had caused a shift in public opinion which helped small banks to get the public behind the vested interest of unit banks and nation-wide deposit insurance: "Had there be no Great Depression, it seems unlikely that the United States would have adopted deposit insurance." (Calomiris & White 1994, p. 146). Consequently, federal deposit insurance was supported by members of Congress who saw deposit insurance as a "bulwark of the unit banking system" (Economides, Hubbard & Palia 1995, p. 22-23).

Although there would have been other options that would have better addressed the banking problem, the legislature, by establishing the FDIC, chose to conserve the weak and overbanked system and thereby to reconfirm the system of unit banking. Williams (1935, p. 105) notes: "Having the kind of banking system that we do, it seems necessary also to have deposit insurance, though it would seem much more desirable to improve the banking system itself than merely to protect the depositor against its defects."

The Banking Act of 1933 thus reconfirmed the branching restriction the McFadden Act had established in 1927. Instead of allowing branching in order to address the banking system instability, federal deposit insurance was adopted, which attempted to address the unstable banking systems' symptoms – banking failures – without pursuing the necessary financial reform to make the banking sector safer. Insofar as even the justification for deposit insurance derived from the hypothesis that illiquidity rather than solvency had caused the large number of bank failures turned out to be incorrect. Calomiris (2007, p. 1) notes: "Pancis or 'contagion' played a small role in bank failure, during or before the Great Depression-era distress. Ironically, the government safety net, which was designed to forestall the (overestimated) risks of contagion, seems to have become the primary source of systemic instability in banking in the current era."

Federal deposit insurance contributed in several ways to the development of TBTF. The attempt to preserve a dysfunctional banking system based on small, predominantly unit banks made the adoption of federal deposit insurance as a perceived remedy almost inevitable. As with prior bank obligation insurance schemes, the creation of the FDIC mainly is attributable to the attempt to preserve a weak, outdated system dominated by unit banks at the cost of a more resilient system based on branch banking. However, instead of stabilising the banking system, the federal safety net made the banking sector even more fragile and thereby set the pre-conditions for TBTF:

(a) By not linking insurance premia to the risk characteristics of banks, federal deposit insurance provided risky banks with a de facto subsidy for risk-taking. As deposit insurance decouples the probability of repayment of deposits from the solvency of banks, deposits were rendered insensitive to the risk level of banks, allowing risky banks to attract more deposits at lower rates than socially optimal (Kane 1989; Hellmann, Murdock & Stiglitz 2000). Various studies confirm that risky banks increase their reliance on insured, and hence less costly, deposits at the expense of non-insured deposits, whose costs responds more effectively to any given bank’s health. Hence, the

37 Bradley (2000) views the introduction of federal deposit insurance before the background of the Great Depression even as “inevitable”. Flood (1992, p. 54-55) points out that in response to the Great Depression, a wide range of perceived solutions had been brought forward, including “relatively sober proposals [...] to scrap the inefficient bureaucracies of representative democracy in favor of a fascist dictatorship or state socialism [...]” and installing a technocratic leadership. “Relative to alternatives such as these, federal deposit insurance [...] was a remarkably moderate option.”

38 For example, Billett, Garfinkel & O’Neal (1998) show that deposit insurance decreases market discipline and
systematically too important to let them fail. As Kovacevich (1996) asks:

“One wonders if the Barings Bank failure would have been treated as a systemic risk problem if it happened to a U.S. bank where regulators have easy access to ‘free’ funds to solve any problem. How would the Drexel, Burnham and Lambert insolvency have been handled if it had been a bank?”

4 Conclusion

This paper developed two points. First, the economic history of the United States’ past two hundred years indicates that financial sector regulation had, from c. 1800 onwards for about 150 years, been dominated by policies that benefitted small banks at the cost of large branching banks. This approach resulted in a fragmented banking sector dominated by small unit banks and characterised by fragility. Ultimately, the collapse of the banking sector during the Great Depression resulted in federal deposit insurance which kept in place the dysfunctional banking system. Not only did federal deposit insurance not address the structural deficits of the banking system, it aggravated the danger emanating from banks by encouraging risk-taking. Deposit insurance allowed banks to maintain riskier balance sheets and grow faster than without the government safety net. Ultimately, the FDIC policy adopted regarding bank failures resulted in the development of the TBTF problem.

Right from the beginning, FDIC policies exhibited a market tendency of bailing out creditors at large banks, while subjecting debtholders at smaller banks to the ordinary resolution process and thus to losses. The inequitable approach to handling banks failures of varying size, in turn, strongly incentivised banks to acquire TBTF size. It is not reading too much into history that in the light of the role the FDIC played in what would commonly be referred to as TBTF, the introduction of federal deposit insurance both precipitated and made possible TBTF. Detrimental to the intention of the public and policy-makers in the 1933, the FDIC played a crucial role in destabilising the banking sector by encouraging inordinate risk-taking by banks.

Second, history shows that a significant share of legislation and bank regulation can be explained by the economic theory of regulation. That is, regulation was supplied in response to the demand of concentrated interest groups at the expense of more dispersed groups which cannot lobby as effectively as the former. While early evidence supporting this notion is mainly anecdotal due to a lack of sufficient data, a wide range of bank regulation since the early 1900s is explained by statistical evidence. However, the most profound regulatory changes can only be explicated by regulatory cognitive capture. In contrast to the pure concept of capture, regulatory cognitive capture allows interest groups to exploit pre-existing resentments, beliefs and ideology of large groups –

Deeply impaired market discipline resulting from deposit insurance allowed risky banks to become TBTF in the first place. Not only did deposit insurance encourage the development of the TBTF doctrine but cross-country evidence indicates that deposit insurance systems increase the propensity of financial and banking crises (Demirgüç-Kunt & Detragiache 2000). Therefore, in addition to encouraging risky behaviour on the part of large banks, deposit insurance fostered an environment in which TBTF banks indeed require TBTF bailouts.

(b) The inequitable approach to handling bank failures adopted by the FDIC by granting large bank creditors full protections in many cases, while creditors of small banks did not receive similarly favourable treatment, represents a major factor for the development of TBTF. The FDIC’s biased policy approach set in motion an autocatalytic process that fed upon itself and provided banks with a non-technical incentive to grow in size in order to receive TBTF protection.

(c) Related to point (b) above, once TBTF institutions exist in numbers, the probability of bailing out uninsured creditors at large banks increases as large bank failures may threaten the solvency of the deposit insurance fund itself. Jones & Oshinsky (2009) estimate that due to the merger activity of the 1990s (which, at least partly, had been motivated by the urge to achieve TBTF size) the probability of Bank Insurance Fund (BIF) insolvency tripled from 1.5% in 1990 to 4.5% in 2005. As “the health of the BIF has become inextricably tied to the health of the top-25 banking organizations and even more so the health of the top-10” (p. 84), a strong incentive for TBTF rescues has been introduced. As the authors further note, the probability of BIF insolvency may be substantially decreased – to 0.3% – if the top-10 institutions are not allowed to fail (p. 84). Therefore, the necessity of preventing the deposit insurance fund from failing lends superficial justification to bailing out large financial institutions, ostensibly for society’s good.

(d) Finally, Stern & Feldman (2009, p. 83) and Kovacevich (1996) point out that pre-funded deposit insurance systems increase the probability of TBTF bailouts by providing policy makers a source of funds that can easily be used to protect uninsured creditors. Therefore, the FDIC-administered Deposit Insurance Fund (DIF) provided the FDIC with an easily accessible source of funds that facilitated ad-hoc bailouts of large banks which were considered to be

...
often the public – by offering a solution that seemingly is in the best interest for the public, while in fact it provides rents to special interest groups. Distrust and outright hostility of the public directed towards big “money centres” historically allowed small financial institutions to lobby successfully for beneficial regulation and legislation in many cases. Often such regulatory capture was augmented by the government’s own interests. Alternatively, capture of legislators by interest groups may function well when external shocks, for example crises, lead to an alignment of public and private interests. As Komai & Richardson (2011) show, a majority of financial crises have been followed by financial sector reformation. However, rarely have these reforms been successful in rendering the financial system more effective. Adoption of deposit insurance is merely one, although particularly disturbing, example for how politicians and the public have been captured by an industry’s vested interests which have led to non-optimal regulation in the wake of financial crises.

Woodward (1998, p. 18) notes: “Anyone who raises her voice as a cynical capture theorist is at heart an advocate for reform”. Although this paper developed a strictly positive explanation for regulation, it possesses some normative value. Given the deep impact of legislation directly following crises, it calls for reforming the financial systems in good times, when the merits of proposed changes introduced by special interest groups may be considered more objectively, appropriately and under less strain to act quickly.

References

23. Mark Carlson & Kris J. Mitchener – Branch Banking as a Device for Discipline: Competition and Bank


27. Board of Governors of the Federal Reserve System, Federal Reserve Committee on Branch, Group, and Chain Banking – Branch Banking in the United States. n.a. 1932.


68. Patrick Slovik – Systemically Important Banks and Capital Regulation Challenges. OECD Economics

69. Department Working Papers 916, 2012. URL: http://dx.doi.org/10.1787/5kg0ps8q8q6-en


