Bank panics and scale economies

David Andolfatto
Federal Reserve Bank of St. Louis
Simon Fraser University

Ed Nosal
Federal Reserve Bank of Chicago

January 2017

Abstract

A *bank panic* is an expectation-driven mass redemption event that results in a self-fulfilling prophecy of losses on demand deposits. We replace sequential service in the Green and Lin (2003) version of the Diamond and Dybvig (1983) model with scale economies in investment opportunities: higher risk-adjusted returns are available for investments that require larger fixed costs. We demonstrate how scale economies and private information are both necessary to generate an equilibrium bank panic. Floating net asset valuation methods—which recent regulations have imposed on some money mutual funds—do *not* eliminate panics. However, restrictions that resemble redemption gates and fees—also imposed on some money mutual funds—can eliminate panics, albeit at the cost of reduced risk-sharing. The expected cost of eliminating panics, however, falls as banks become larger or more interconnected. Finally, we discuss how our theory is consistent with the notion that a low-interest rate policy induces a reach-for-yield behavior that potentially manifests itself as a less stable financial system.

1 Introduction

Investors in market economies have access to an array of financial products with implicit or explicit options exercisable at their discretion. Bryant (1980) suggests that options like demandable and callable debt are a means to insure investors against unobservable risks. The same could be true of any short-term debt instrument, like commercial paper and repo, since rollover is an option.

The benefit of insurance in this form, however, is undermined if options are triggered *en masse* by fear instead of fundamentals. Diamond and Dybvig (1983) were the first to show how demand deposit liabilities could result in a *bank panic*—an episode where the mere expectation of large withdrawals of funding from an otherwise solvent banking system could result in a self-fulfilling prophecy of insolvency.

While no one questions the emotional trauma associated with financial crises, the role that psychological factors play in triggering such events remains an open question. An alternative view is that financial instability is merely symptomatic of difficulties experienced in the broader economy during an economic downturn. For example, Gorton (1988) notes that banking crises during the U.S. National Banking era were frequently (though not always) preceded by economic recession. There is surely some merit to this view. But while this timing is suggestive, it is not conclusive. Moreover, even if it is true in general, it appears not to be true in all cases; see also Capiro and Klingebiel (1997), who report that some of the (over 80) banking crises they document are not associated with recessions. Hence, the question of whether bank *panics* are possible remains salient. This is especially so for policymakers, whose interventions must be guided with a view of the ultimate sources of financial instability.

Central to the *theory* of bank panics in the tradition of Diamond and Dybvig (1983) is the notion of *sequential service*—the practice of satisfying depositor withdrawal demands on a first-come-first-served basis. However, sequential service is not sufficient to guarantee bank panics even in the presence of aggregate liquidity risk and private information over liquidity needs (Green and Lin 2000, 2003). Moreover, as emphasized by Wallace (1988), the logic of sequential service seems inconsistent with deposit insurance as a remedy for bank panics. Our own view is that while sequential service may have some role to play in panic episodes, there may exist other, complementary sources of instability. We suggest below that panics can arise as a by-product of the interaction of scale economies in investments and the demandable liability structures that finance them.

Central to our idea is the notion that investments generally entail a fixed cost and that, in general, the larger the fixed cost the larger the return. Some investments, like cash or bonds, entail a small fixed cost and promise a low rate of return. These investments can be viewed as offering close to a linear return schedule. Other investments, like commercial real estate, offer a higher rate of return, but only if funded at scale. Our assumption is that the risk-adjusted rate of return from funding (say) a large shopping mall dominates the rate of return afforded from funding (say) a small office unit. We view the assumption of scale economies as a self-evident fact—almost any operation surely entails some fixed cost. Our scale assumption is consistent with the fact that, empirically, large firms are more productive, or pay more, than small firms. It is also consistent with an empirical literature that documents the existence of scale economies in banking.[1]

Our analysis is based on the Green and Lin (2000, 2003) finite-trader version

[1] See, for example, Hughes and Mester (2013), Beccali, Anolli, and Borello (2015), and Wheelock and Wilson (2016).

of the Diamond and Dybvig (1983).[2] The standard formulation assumes a constant returns to scale investment technology and a sequential service constraint. We assume that the investment technology is characterized by increasing returns and that there is no sequential service. As is standard, we assume that all agents in the model have access to the same investment technology and that the distinguishing characteristic of banks is how these investments are funded. In particular, banks offer their depositor base products that provide the flexibility to withdraw funding on short notice.

We demonstrate that neither private information nor scale economies can *in themselves* produce a bank panic—the *ex ante* efficient allocation is uniquely implementable when one or the other of these properties is present. However, when these properties coexist, the efficient allocation is only weakly implementable. In this case, a bank panic equilibrium exists. Panic equilibria can be eliminated by requiring some long term borrowing (or requiring a minimum level of certain assets holdings), but only at the cost of reduced risk-sharing. However, the expected cost of eliminating the bank panic equilibrium is decreasing in bank size. Larger (or more connected) banking systems can be made more stable at a lower cost–a prediction that bears up well against the historical evidence (Williamson, 1989).

2 The model

The model is a finite-trader version of the Diamond and Dybvig (1983) model with aggregate liquidity risk as developed by Green and Lin (2000, 2003). The economy has two dates, $t = 1, 2$, and a finite number $N \geq 3$ of *ex ante* identical agents. Agents are subject to a shock at date $t = 1$ that determines their preference type: *impatient* or *patient*. Let $0 < \pi < 1$ denote the probability that an agent becomes impatient. Let π_n denote the probability that $0 \leq n \leq N$ agents are impatient. We assume that agent types are *i.i.d.* so that $\pi_n = \binom{N}{n}\pi^n(1-\pi)^{N-n}$. Note that $0 < \pi_n < 1$ for all n (the distribution of types has full support).

Impatient agents want to consume at date 1 only. Patient agents are willing to defer consumption to date 2; technically, they are indifferent between consuming at dates 1 and 2. Let c_t represent the consumption of an agent at date t. *Ex ante* preferences are given by

$$U(c_1, c_2) = \begin{cases} u(c_1) & \text{w.p.} \quad \pi \\ u(c_1 + c_2) & \text{w.p.} \quad 1 - \pi \end{cases}, \tag{1}$$

where $u(c) = c^{1-\sigma}/(1-\sigma)$, $\sigma > 1$.

Each agent is endowed with a claim to y units of date 1 output. There is a technology that transforms k units of date 1 output into $F_\kappa(k)$ units of date 2

[2] For a comprehensive review of this literature, see Ennis and Keister (2010).

output according to

$$F_\kappa(k) = \begin{cases} rk & \text{if} \quad k < \kappa \\ Rk & \text{if} \quad k \geq \kappa \end{cases}, \tag{2}$$

where $0 < r < 1 < R$ and $0 \leq \kappa < Ny$. The high rate of return R is available only if the level of investment exceeds a minimum scale requirement of κ.[3] When the minimum scale is not met the rate of return is a cost to storage, indexed by the parameter r (the storage cost is $1-r$). The function (2) is a generalization of the standard technology used in this literature, where the standard technology has $\kappa = 0$ so that $F_0(k) = Rk$ for all $k > 0$.

The benefits of cooperation in this economy are twofold. First, there are the usual gains associated with sharing risk. Second, and absent from the standard model, scale economies are more easily attainable when resources are pooled. For convenience, we relabel agents as *depositors* and call any contractual arrangement among depositors to share risk and exploit scale economies a *bank*. Formally, a bank is a resource-allocation mechanism that pools the resources of $0 \leq \hat{N} \leq N$ depositors—before they learn their types—in exchange for a structured liability product that delivers consumption (or withdrawal limits) at dates 1 and 2, conditional on individual and aggregate information available at each date.

An important consideration in the design of the contract is the nature of private information and communications. What makes our environment a bank problem—as opposed to a standard insurance problem—is that the demand for early consumption—date 1 withdrawals—is based on private information. Consequently, the contract must embed an option that is exercisable at the depositor's discretion at the early date. In this sense, the bank's contract resembles a standard demand deposit liability. Moreover, for the purpose of ensuring efficient resource allocation, the contract should be structured in a way that gives depositors an incentive to represent their liquidity preferences truthfully. That is, the allocation implied by the contract should be *incentive compatible*.

Following their decision to participate in a banking arrangement, we assume that depositors can communicate with the bank *either* at date 1 *or* at date 2, but not at both dates. This permits us to model the bank as a simple direct revelation mechanism along the lines of Peck and Shell (2003). That is, a depositor implicitly communicates "I am impatient" by visiting the bank at date 1 and requesting a withdrawal. Likewise, a depositor implicitly communicates "I am patient" by postponing the visit to a later date. One virtue of modeling a depositor's "travel itinerary" in this manner is that it accords well with actual depositor behavior. In particular, depositors are not typically in the habit of making regular calls to their bank to inform the credit manager that they will not be needing any funds today. Having said this, our communication technology *may* be restrictive in some circumstances. Just as sequential service

[3] One could easily generalize the analysis to permit multiple threshold levels and associated rates of return, but here we stick to one threshold level for the sake of making our point in the simplest manner possible.

imposes restrictions on allocations, our communications technology—which permits depositors to visit the bank at either date 1 or date 2 but not both—may also adversely impinge on implementable allocations. We will identify in due course when the communication technology does and does not place additional restrictions on allocations.

Given our specification of depositor travel itineraries, consumption limits at each date need only be conditioned on the number of agents m visiting the bank at date 1, where $m \in \{0, 1, ..., \hat{N}\}$. In particular, if m agents visit the bank at date 1, then each agent receives $c_1(m)$ units of date 1 consumption. Agents who visit the bank at date 2 each receive $c_2(m) = F_\kappa[\hat{N}y - mc_1(m)]/(\hat{N} - m)$ units of date 2 consumption. Hence, the bank offers depositors an allocation or *contract* $(\mathbf{c}_1, \mathbf{c}_2)$, where $\mathbf{c}_1 = [c_1(1), \ldots, c_1(\hat{N})]$ and $\mathbf{c}_2 = [c_2(0), c_2(1), \ldots, c_2(\hat{N} - 1)]$.

In what follows, we restrict attention to equilibria in which $\hat{N} = N$. It seems clear enough that whenever a scale economy is present, $\hat{N} = 0$ can be an equilibrium outcome.[4] But it is an outcome that is attributable solely to the assumed non-convexity in production and nothing to do with how investment is financed. Our concern in what follows is to examine the potential indeterminacy introduced by the funding structure itself when N agents deposit their endowment at the bank.

3 Efficient incentive-compatible allocations

We characterize the properties of efficient incentive-compatible allocations. We begin with the standard linear case where the return to investment is invariant to its scale, $\kappa = 0$. We then study the case where the return to investment is increasing in scale, $\kappa > 0$.

3.1 Linear technology

Assume, for the moment, depositor types are observable. Since $\kappa = 0$, efficiency dictates that impatient depositors visit the bank at date 1 and patient depositors visit the bank at date 2. Clearly, impatient depositors strictly prefer to visit the bank at date 1 when $c_1(n) > 0$. Patient depositors may not want to wait to visit until date 2, but since we assume their type is observable, they can always be punished for arriving early (by denying them service, for example). Thus, when depositor types are observable, we have $m = n$. The efficient allocation

[4] When $\kappa > y$, a depositor's payoff associated with an equilibrium where $\hat{N} = 0$ is $u(y)$, i.e., all depositors consume their endowments at date 1. Suppose a depositor defects from proposed equilibrium play and deposits him claim at the bank. In this situation, the optimal bank contract would require that the depositor—independent of his type—visit the bank at date 1 and receive y. Hence, a depositor does not strictly prefer to depositor his endowment at the bank and, as a result, $\hat{N} = 0$ is an equilibrium.

depends on the number of depositors n that visit the bank at date 1.[5]

When $\kappa = 0$ and types are observable, the bank chooses a deposit contract $\{c_1(n)\}$ that maximizes the expected utility of an (*ex ante* identical) representative depositor,

$$\max_{\{c_1(n)\}} \sum_{n=0}^{N} \pi_n \{nu\,[c_1(n)] + (N-n)u\,[c_2(n)]\} \tag{3}$$

subject to the resource constraints

$$\begin{aligned} nc_1(n) &= Ny - k(n) && (4)\\ Rk(n) &= (N-n)c_2(n) && (5) \end{aligned}$$

which, when combined, yields

$$nc_1(n) + \frac{(N-n)c_2(n)}{R} = Ny, \tag{6}$$

for all $n \in \{0, 1, ..., N\}$. Let $(\mathbf{c}_1^*, \mathbf{c}_2^*)$ denote the solution to the problem above. It is easy to see that there is a unique solution that satisfies

$$u'[c_1^*(n)] = Ru'[c_2^*(n)] \ \forall n, \tag{7}$$

and the resource constraint (6). Given our CES preference specification, the closed-form solution is

$$\begin{aligned} c_1^*(n) &= \frac{Ny}{n + (N-n)R^{1/\sigma - 1}} && (8)\\ c_2^*(n) &= R^{1/\sigma} c_1^*(n). && (9) \end{aligned}$$

Notice that $\sigma > 1$ implies $R^{1/\sigma-1} < 1$ since $R > 1$. It follows then that both $c_1^*(n)$ and $c_2^*(n)$ are decreasing in n. As in Wallace (1988) one can interpret $c_1^*(n) > c_1^*(n+1)$ as a partial suspension of payments in the sense that the bank suspends payments to each depositor by $c_1^*(n) - c_1^*(n+1)$ if an additional depositor shows up at date 1. By construction, these partial suspensions are efficient. Using (5) and (8), the associated efficient investment or capital schedule, $k^*(n)$, is

$$k^*(n) = \frac{(N-n)R^{1/\sigma-1}Ny}{n + (N-n)R^{1/\sigma-1}}. \tag{10}$$

Notice that $k^*(n)$ is decreasing n and that $k^*(n) < (N-n)y$.[6]

A large value for n means that the aggregate demand for early withdrawals is high. In this case, it makes sense to devote less resources to investment, reducing

[5] Green and Lin (2000, 2003) also provide a characterization of the efficient allocation when there is no sequential service and the investment technology is linear.

[6] The latter inequality is a direct implication of risk sharing. Capital per patient depositor, $k^*(n)/(N-n)$ is also a decreasing function of n.

the effective return for late withdrawals, spreading the additional early resources more thinly among the more numerous impatient depositors. Note that high realizations for n can be interpreted as recessionary events or investment collapses associated with large numbers of depositors making early withdrawals. These events, however, are driven by economic fundamentals. A bank (or the banking sector) could mitigate the economic impact of these "fundamental runs" by expanding its depositor base, i.e., increasing N. Our full support assumption, however, implies that the probability that all depositors desire early withdrawal, π_N, will remain strictly positive, even if $\pi_N \to 0$ as $N \to \infty$.

3.1.1 Private information

Assume now that liquidity preferences are private information. Depositors may misrepresent themselves for private gain at the expense of the community. Since depositor behavior will generally depend on how they believe other depositors are likely to behave, private information renders the environment strategic.

A *depositor game* is played after agents deposit their date 1 endowment claims at the bank. Specifically, after they learn their types, depositor $j \in \{1, 2, ..., N\}$ chooses an action $t_j \in \{1, 2\}$, where t_j denotes the date depositor j visits the bank. Depositor j knows only his own type when he chooses t_j and, e.g., he does not know the number of impatient depositors n in the economy. A *strategy profile* $\mathbf{t} \equiv \{t_1, t_2, ..., t_N\}$ implies a number $m \in \{0, 1, ..., N\}$, the number of depositors that visit the bank at date 1. A *truth-telling strategy* is a strategy profile that has impatient depositors visiting the bank at date 1 and patient depositors visiting the bank at date 2. Hence, a truth-telling strategy implies that $m = n$. A *panic strategy* is a strategy profile that has all depositors visiting the bank at date 1, i.e., $t_j = 1$ for all j and, as a result, $m = N$ for any $n < N$.

A strategy profile $\mathbf{t}$ (and its associated m) constitutes a Bayes-Nash *equilibrium* of the depositor game with deposit contract or allocation $(\mathbf{c}_1, \mathbf{c}_2)$ if $t_j \in \mathbf{t}$ is a best-response for depositor j against $\mathbf{t}_{-j} \equiv \{t_1, ..., t_{j-1,} t_{j+1}, ..., t_N\}$ for all $j \in \{0, 1, ..., N\}$. An allocation $(\mathbf{c}_1, \mathbf{c}_2)$ is said to be *incentive-compatible* (IC) if a truth-telling strategy is an equilibrium for the depositor game. Since, in any equilibrium $c_1(m) > 0$, impatient depositors always tell the truth—they visit the bank at date 1. A patient depositor tells the truth by visiting the bank at date 2; he has an incentive to visit the bank at date 2 iff

$$\sum_{n=0}^{N-1} \Pi^n u\left[c_2(n)\right] \geq \sum_{n=0}^{N-1} \Pi^n u\left[c_1(n+1)\right], \tag{11}$$

where Π^n is the conditional probability that there are n impatient agents given there is at least one patient agent, i.e.,

$$\Pi^n = \frac{\binom{N-1}{n}(1-\pi)^{N-n-1}\pi^n}{\sum_{n=0}^{N-1}\binom{N}{n}(1-\pi)^{N-n-1}\pi^n}.$$

There are two important results associated with the solution $(\mathbf{c}_1^*, \mathbf{c}_2^*)$. First, it is incentive-compatible. To see this, note that when $R > 1$, (8) and (9) imply that $c_2^*(n) > c_1^*(n) > c_1^*(n+1)$. In words, a patient depositor's consumption is higher than an impatient depositor's consumption in a truth-telling equilibrium. And, if a patient depositor chooses to visit the bank at date 1, then date 1 consumption for all depositors is lower compared to what they would enjoy if the patient depositor instead visited the bank at date 2. Therefore, assuming that all other depositors are playing truthfully, a patient depositor has no incentive to visit the bank at date 1 since doing so would result in a strictly lower payoff.

Second, the truth-telling equilibrium that implements $(\mathbf{c}_1^*, \mathbf{c}_2^*)$ in the depositor game is *unique*. To see this, first note that it is a dominant strategy for impatient depositors to visit the bank at date 1 since $c_1(m) > 0$ for all $m \in \{1, 2, ..., N\}$. It is also a dominant strategy for the patient depositor to visit the bank at date 2. To understand this, let the patient depositor conjecture *anything* about the behavior of other depositors. Any such conjecture generates an $m \in \{1, 2, ..., N\}$ for that patient depositor. If the patient depositor plays truthfully, he gets $c_2^*(m) > c_1^*(m) > c_1^*(m+1)$, where these inequalities follow from the properties of the allocation (8) and (9) when $R > 1$. Hence, regardless of the volume of early withdrawals, a patient depositor will always get a higher payoff by leaving his resources with the bank at date 1 (and letting the deposit earn a high rate of return) than visiting the bank to withdraw funds at date 1, i.e., $c_2^*(m) > c_1^*(m+1)$. Hence, it is a dominant strategy for a patient agent to visit the bank at date 2 and, as a result, allocation $(\mathbf{c}_1^*, \mathbf{c}_2^*)$ is uniquely implemented (in dominant strategies).

Notice that our communication technology—or travel itinerary protocol—is not a binding restriction when the investment technology is linear. This is because when depositor types are private information, the first-best allocation $(\mathbf{c}_1^*, \mathbf{c}_2^*)$ is uniquely implementable under our communication technology (that depositors can visit the bank at either date 1 or date 2 but not at both dates). This implies that private information *by itself* cannot be responsible for bank panics in this environment.

Green and Lin (2003) and Peck and Shell (2003) implicitly examine if unique implementation survives sequential service.[7] We take an alternative approach and examine if scale economies in investment technology impairs unique implementation.

3.2 Scale economies

Consider an economy with $0 < \kappa < Ny$ (recall that $N \geq 3$). As above, we begin by characterizing the efficient allocation assuming that preference types

[7] Green and Lin (2003) obtain unique implementation even after sequential service is imposed; for some model parameterizations Peck and Shell (2003) are able to generate multiple equilibrium—a truthtelling equilibrium and a panic equilibrium—when there is sequential service.

are observable and depositors are subject to our communication technology (or travel itinerary protocol). Let $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ denote the efficient truth-telling allocation, where the "hat" indicates that scale economies are present. We will later verify that this allocation is also incentive-compatible when depositor types are private information.

To demonstrate our insights and results in the cleanest possible way, we assume that $\kappa = 2y$.[8] It follows that $k^*(N-2) < 2y < k^*(N-3)$, where $k^*(n)$ is described by (10). In other words, in the event that there are $n \leq N-3$ impatient depositors—or, equivalently, in the event that there are at least three patient depositors—the first-best level of investment exceeds the minimum scale κ. It follows immediately that $\{\hat{c}_1(n), \hat{c}_2(n)\} = \{c_1^*(n), c_2^*(n)\}$ for $n \in \{0, 1, ..., N-3\}$. Moreover, when all depositors are patient, we have that $\hat{c}_1(N) = y = c_1^*(N)$, i.e., all depositors retrieve their initial deposit y at date 1. Consequently, we have

Property 1 $[\hat{c}_1(n), \hat{c}_2(n)] = [c_1^*(n), c_2^*(n)]$ for all $n \in \{0, 1, ..., N-3, N\}$.

We now characterize the efficient allocation for the remaining two cases: $n = N-1$ and $n = N-2$. Since $k^*(N-j) < \kappa$, $j = 1, 2$, $\hat{k} = k^*(N-j)$ is not a solution to the bank's problem. The solution will take one of two forms involving either the high-return option, R, or the low-return option, r. That is, it will be optimal to set to $\hat{k} = \kappa$ and achieve the high rate of return R, or to choose $\hat{k} < \kappa$ and get the how rate of return $r < 1$. Before proceeding with the analysis, we provide some insight into why depositors may want to enter into a risk-sharing arrangement that is subject to the low rate of return $r < 1$ in some states of the world.

To begin, instead of subjecting depositors to the low return $r < 1$, why not simply return the initial deposit y to all depositors at date 1 in these states? The answer lies in our assumed communication friction: depositors can only visit the bank at either date 1 or date 2 but not at both dates. Because depositors only know their own type and not the state of the world n when making their withdrawal (travel) decision, the risk-sharing mechanism can do no better than have patient depositors (unconditionally) travel at date 2. Of course, incentive-compatibility of the constrained efficient allocation ensures that patient depositors prefer to delay their withdrawal understanding that, *ex post*, their realized rate of return may turn out to be low in some states of the world (i.e., when $n = N-1$ and/or $n = N-2$).

Second, why does the bank contract not ensure that the level of investment is always greater than or equal to κ, so that there are no states in which the low rate of return $r < 1$ prevails? Note that in order to fund investment $k = \kappa$, it may be the case that in some states, e.g., $n = N-1$, consumption of impatient agents is less than y and consumption of patient agents is greater than Ry. Such an allocation has bad risk sharing properties. Depositors (*ex ante*) may prefer

[8] We show below that this assumption is without loss of generality.

an allocation that provides a lower rate of return but has better risk-sharing properties.

Let's now return to examine the remaining two cases: $n = N - 1$ and $n = N - 2$. We begin with the case of two patient depositors, $n = N - 2$. Consider first the high return option, which requires $\hat{k} = \kappa = 2y$. Using (4), we see that $\hat{k} = 2y$ implies $\hat{c}_1^R(N-2) = y$. From (5), we then have $\hat{c}_2^R(N-2) = Ry$. Consider next the good risk-sharing but low return option, which implies that $\hat{k} < \kappa$. In this case, $\hat{c}_1^r(N-2)$ and $\hat{c}_2^r(N-2)$ are determined by replacing R with r in (8) and (9), respectively, and setting $n = N - 2$, i.e.,

$$\hat{c}_1^r(N-2) = \frac{N}{N-2+2r^{1/\sigma-1}}y \tag{12}$$

$$\hat{c}_2^r(N-2) = r^{1/\sigma}\hat{c}_1^r(N-2) \tag{13}$$

Since $\sigma > 1 > r$, it follows that $\hat{c}_2^r(N-2) < \hat{c}_1^r(N-2) < y$. We conclude that the high return option is optimal because it dominates the efficient risk-sharing low-return option, i.e., $\hat{c}_1^R(N-2) > \hat{c}_1^r(N-2)$ and $\hat{c}_2^R(N-2) > \hat{c}_1^r(N-2)$. Hence, we have

Property 2 When $n = N - 2$, the efficient allocation is given by $\hat{c}_1(N-2) = \hat{c}_1^R(N-2) = y$ and $\hat{c}_2(N-2) = \hat{c}_2^R(N-2) = Ry$.

We now examine the remaining case, $n = N - 1$. Once again we need to evaluate the relative payoffs associated with the high and low return options. To achieve the high return, the level of investment must be set to $\hat{k} = \kappa = 2y$. Using (4) and (5), we see that $\hat{k} = 2y$ implies $\hat{c}_1^R(N-1) = y(N-2)/(N-1)$ and $\hat{c}_2^R(N-2) = 2Ry$. Since $\hat{c}_1^R(N-1) < y < R < \hat{c}_2^R(N-2)$, the high return option comes at the price of reduced risk-sharing.

Consider next the low return option, which implies that $\hat{k} < \kappa$. In this case, $\hat{c}_1^r(N-1)$ and $\hat{c}_2^r(N-1)$ are determined by replacing R with r in (8) and (9), respectively, and setting $n = N - 1$, i.e.,

$$c_1^r(N-1) = \frac{N}{N-1+r^{1/\sigma-1}}y \tag{14}$$

$$c_2^r(N-1) = r^{1/\sigma}c_1^r(N-1) \tag{15}$$

It is straightforward to see

Lemma 1 *For* r arbitrarily close to (but less than) unity, $c_1^r(N-1) \approx \hat{c}_2(N-1) \approx y = \hat{c}_1(N)$, where $c_1^r(N-1) < \hat{c}_2(N-1) < y$.

Thus, for r sufficiently close to unity, the expected utility payoff to taking the low return option is approximately $Nu(y)$, while the expected utility associated with the high return option is

$$(N-1)u\left(\frac{N-2}{N-1}y\right) + u\left(2Ry\right).$$

If r is sufficiently close to unity and if $\sigma \geq 2$, then the expected utility associated with the low return option exceeds that of the high return option, i.e.,

$$(N-1)u\left(\frac{N-2}{N-1}y\right)+u\left(2Ry\right)<Nu\left(y\right).$$

To see this, note that our functional form for $u(\cdot)$ implies that the above inequality becomes

$$(N-1)(1-\sigma)^{-1}\left(\frac{N-2}{N-1}y\right)^{1-\sigma}+(1-\sigma)^{-1}\left(2Ry\right)^{1-\sigma}<N(1-\sigma)^{-1}(y)^{1-\sigma}.$$

Since $1-\sigma<0$, the above inequality becomes

$$(N-1)\left(\frac{N-2}{N-1}\right)^{1-\sigma}+\left(2R\right)^{1-\sigma}>N. \tag{16}$$

Since $\sigma \geq 2$, clearly

$$\left(\frac{N-1}{N-2}\right)^{\sigma-1}>\frac{N}{N-1} \tag{17}$$

is a valid inequality because $(N-1)/(N-2)>N/(N-1)$. Since condition (17) holds, then so does condition (16). Therefore, we have

Property 3 When $n=N-1$, $r<1$ sufficiently close to unity and $\sigma \geq 2$, the efficient allocation is given by $\hat{c}_1\left(N-1\right)=\hat{c}_1^r\left(N-1\right)$ and $\hat{c}_2\left(N-1\right)=\hat{c}_2^r(N-1)$ as described by (14) and (15), respectively.

Property 3 has the feature that when $N-1$ the bank "breaks the buck" in the sense that for every unit that agents deposit at the bank, they receive less than a unit payoff at date 1 (and date 2). The relevance of this observation is explained in Section 5, below.

Properties 1-3 allow us to characterize the efficient allocation when $\kappa=2y$, $r<1$ sufficiently close to unity and $\sigma \geq 2$. In particular, the efficient allocation, $(\hat{\mathbf{c}}_1,\hat{\mathbf{c}}_2)$, is given by

$$\begin{aligned}(\hat{\mathbf{c}}_1,\hat{\mathbf{c}}_2)=\{[c_1^*(n),c_2^*(n)]_{n=0}^{N-3},&\left[\hat{c}_1^R(N-2),\hat{c}_2^R(N-2)\right],\\ &\left[\hat{c}_1^r(N-1),\hat{c}_2^r(N-1)\right],\left[c_1^*(N),c_2^*(N)\right]\}\end{aligned} \tag{18}$$

A few remarks are in order.

1. The allocation $(\hat{\mathbf{c}}_1,\hat{\mathbf{c}}_2)$ respects the assumed communication technology, namely, that depositors can only visit the bank at date 1 *or* date 2 *but not both dates*. The communication technology is restrictive in some states of the world: if depositors could visit the bank at dates 1 and 2, the bank would provide a payment y to *all* depositors at date 1 in state $n=N-1$ instead of $\hat{c}_2\left(N-1\right)<\hat{c}_1\left(N-1\right)<y$.

2. Since $r < 1$, it is straightforward to show that the expected utility of allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ exceeds autarky, where $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ has impatient depositors visiting the bank at date 1 and patient depositors visiting at date 2.[9]

3. We assumed that $\kappa = 2y$. In fact, the qualitative properties of the solution remain intact for more general cases. In particular, consider an arbitrary minimum scale $\kappa > 2y$. Then, there will be some $3 < j \leq N$ for which Property 1 above will continue to hold for $n \in \{0, 1, ..., N-j, N\}$. The remaining state contingent allocations $N - j < n < N$ are then characterized either by Property 2 or 3. For minimum scales $\kappa > 2y$, the set of states, n, for which Property 3 holds simply becomes larger. Our main result depends on there being at least one such state, for which $\kappa = 2y$ turned out to be sufficient for this purpose.

Up to this point we have assumed that depositor types are observable to the bank. An implication of this assumption is that allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ can be implemented uniquely if the bank threatens to punish any patient depositor arriving at date 1.[10] Thus, the scale in investment on its own in a public information economy cannot generate bank panics, just as private information on its own in a linear economy cannot. We now re-introduce private information regarding depositors' types when scale economies are present.

3.2.1 Private information

We now assume that depositor types are private information. Our first proposition asserts that the efficient allocation is implementable as a truth-telling equilibrium.

Proposition 1 *The allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ described by (18) can be implemented as a truth-telling equilibrium, with impatient depositors choosing to visit the bank at date 1 and patient depositors choosing to visit the bank at date 2.*

The proof of this proposition is straightforward. First, for any number of impatient depositors $n \leq N-2$, $\hat{c}_2(m) > \hat{c}_1(m) > \hat{c}_1(m+1)$. This implies that a patient depositor will be *strictly* better off visiting the bank at date 2—as

[9] In autarky has all agents consume their endowment at date 1. Notice that autarky is equivalent to banking where *all* depositors visit the bank at date 1 and no depositors visiting at date 2.

[10] Since depositor type is observable, depositors visiting the bank at date 1 automatically reveal their type. A patient depositor visiting at date 1 is not supposed to be there, so the bank can refuse to service him (in accordance to the contractual terms which would have specified this denial-of-service stipulation *ex ante*). Assuming that the mechanism can commit to threats it makes along off-equilibrium paths, no patient depositor would ever have an incentive to misrepresent himself since attempting to do so—by visiting at date 1—results in a payoff of zero, which is strictly dominated by visiting at date 2. Therefore, in the absence of private information, and conditional on participation, $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ is uniquely implementable as an equilibrium.

opposed to date 1—when at least one other depositor visits the bank at date 2. When the patient depositor is the only only patient type in the population, he would receive a higher payoff if he visits the bank at date 1. But, of course, the depositor does not know he is the only patient depositor. Since, however, his date 2 visit payoff $\hat{c}_2(N-1) \approx \hat{c}_1(N) \approx y$ when $r < 1$ is sufficiently close to unity (see Lemma 1), allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ satisfies the compatibility constraint (11).[11] As a result, allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ can be implemented as a truthful equilibrium: patient depositors will visit the bank at date 2 assuming that all of depositors behave truthfully. We now examine if the truthful equilibrium is the unique equilibrium.

4 Panic equilibria

Before agents become depositors, they make the usual participation decision: should they join the risk-sharing arrangement or not? Collectively, this participation decision determines $0 \leq \hat{N} \leq N$. Autarky is always a feasible option and so places a lower bound on *ex ante* utility. Assuming that no individual agent possesses sufficient resources to operate investment at scale, i.e., $\kappa > y$, the autarkic payoff is $u(y)$. Autarky is always an equilibrium in the participation game when $\kappa > y$. That is, if an agent believes that all other agents are choosing not to participate in the banking arrangement, then it is optimal for an agent to follow the crowd. However, this is not the indeterminacy we are interested in. In what follows, we assume depositors coordinate on the participation outcome.

Our main result is stated in the following proposition, which assumes the parameterization above (but which holds more generally).

Proposition 2 *If depositor types are private information, then allocation* $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ *admits a panic equilibrium, where all* N *agents exercise their option to withdraw funding from the bank at date 1, regardless of their true liquidity needs.*

Note that because depositor types are not observable, the bank cannot threaten to punish any patient depositor that arrives early. Moreover, because the distribution of depositor types has full support, suspension clauses are infeasible.[12]

To check the validity of Proposition 2, we begin by proposing a strategy profile in which all N depositors visit the bank at date 1. The question is whether a patient depositor has an incentive to follow the crowd, or leave his deposit in the bank with the intent of withdrawing at the later date. If he plays the proposed panic strategy, he receives a payoff $\hat{c}_1(N) = y$. If he instead

[11] Since $\hat{c}_2(m) > \hat{c}_1(m) > \hat{c}_1(m+1)$ when $n \leq N-2$, $r < 1$ need *not* be sufficiently close to unity to have allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ be incentive compatible.

[12] At least, this is true for direct mechanisms. Andolfatto, Nosal and Sultanum (2016) demonstrate that a broader class of indirect mechanisms could nevertheless be designed to guarantee unique implementation even under the full support assumption.

defects from the panic strategy and travels to the bank at date 2, by Lemma 1, he receives $\hat{c}_2(N-1) \approx y$ where $\hat{c}_2(N-1) < y$. Consequently, a patient depositor is strictly motivated to panic if he conjectures that other patient depositors are as well.[13]

4.1 Eliminating panics

Following Peck and Shell (2003) we could construct sunspot equilibria by introducing a parameter θ, the probability of a sunspot—an extrinsic event observed prior to the investment decision that coordinates depositors to play the panic equilibrium. (With probability $1-\theta$ the sunspot is not observed and depositors play the truthtelling equilibrium.) If we were to do so, the allocation and depositor expected utility would then depend on θ. For example, if contract $(\hat{c}_1, \hat{c}_2)$ is in place and if $Eu(\hat{c}_1, \hat{c}_2)$ represents the expected utility associated with contract $(\hat{c}_1, \hat{c}_2)$ when truthtelling strategies are played, then the expected utility associated with contract $(\hat{c}_1, \hat{c}_2)$ for the representative depositor is $(1-\theta)Eu(\hat{c}_1, \hat{c}_2)+\theta u(y)$. Notice that for higher values of θ, the expected payoff to the risk-sharing arrangement diminishes. However, unlike Peck and Shell (2003), autarky is never preferred to a banking arrangement with the efficient contract $(\hat{c}_1, \hat{c}_2)$ since $(1-\theta)Eu(\hat{c}_1, \hat{c}_2) + \theta u(y) > u(y)$ for all $\theta > 0$.

Similar to Peck and Shell (2003), for θ sufficiently large, the bank may want to eliminate panics by offering a contract that is *panic proof*. The most obvious way to render the risk-sharing arrangement panic proof while maximizing the flexibility afforded by the demandable liability structure is for the bank to invest in at least κ units of capital in all states $m < N$. (In state $m = N$, there are no reported patient depositors so the bank will return all of its Ny potential investment assets.) When $\kappa = 2y$, this implies that the best panic-proof contract, $(\widetilde{c}_1, \widetilde{c}_2)$, is $(\hat{c}_1, \hat{c}_2)$ for all $m \neq N-1$ and, when $m = N-1$,

$$[\widetilde{c}_1(N-1), \widetilde{c}_2(N-1)] = [y(N-1)/(N-1), 2Ry].$$

Because $\widetilde{c}_1(N-1) < \hat{c}_1(N-1) < y < \hat{c}_1(N-2)$, this modification to the original bank contract can be interpretated as an extra haircut that is applied to early redemptions in the high redemption state $m = N-1$. This contract is panic proof because if a sunspot is observed and a patient depositor believes that all other depositors will visit the bank at date 1, he has no incentive to visit the bank early since $\widetilde{c}_2(N-1) = 2Ry > y = \hat{c}_1(N)$. That is, the patient depositor knows that the bank is committed to funding its operations at (at least minimum) scale.

Eliminating bank panics in this manner, however, necessarily implies less risk-sharing, since the allocation in state $N-1$ is inefficient *ex post*. However, the expected cost of eliminating panics in this manner is a decreasing function of N, since the probability π_{N-1} is strictly decreasing in the size of the depositor

[13] If $r = 1$, a patient depositor would be indifferent between misrepresenting himself or not.

base N. Hence, larger banks (or a more interconnected banking system) can be made more stable at a lower expected cost. Thus, the constrained-efficient risk-sharing arrangement may entail a panic proof contract if the probability of a sunspot is sufficiently high and/or if the banking system has a sufficiently large depositor base.

5 Discussion

Our analysis suggests that bank panics can exist even in the absence of sequential service. The essential ingredients are private information over liquidity needs and a desire to fund investments subject to scale economies with demandable (or short-term) debt. Our model implies that organizations that make use of bank credit lines or commercial paper to fund working capital and inventory could suddenly see the value of their operations decline if funding is suddenly pulled, reinforcing a rationally bleak outlook on the backing of their remaining debt. One such sudden event occurred on September 16, 2008, when the Reserve Primary Fund "broke the buck." This news triggered a large wave of redemptions from the money market sector that stopped only after the U.S. government announced it would insure deposits in money market funds.[14] At the time, mutual funds allowed their depositors to withdraw their investments on demand with impunity at a fixed (par) exchange rate.

Note that the mechanism we study permits—indeed, requires—net-asset-valuation (NAV) methods for pricing deposits. In particular, early withdrawal amounts are made conditional on the period's redemption activity and the fund is permitted to break the buck in very heavy redemption states. On July 23, 2014, the SEC announced money market reforms that included the requirement of floating NAV for institutional money market funds.[15] Our model suggests that NAV pricing of demandable liabilities alone is not sufficient to prevent panics. However, note that the SEC reform also stipulated the use of liquidity fees and redemption gates to be administered in periods of stress. These contractual stipulations are consistent with eliminating panics in our model economy. For example, the difference $\hat{c}_1(N-1) - \widetilde{c}_1(N-1) > 0$ can be interpreted as a liquidity fee that depositors pay to obtain funds when redemption activity is judged by the directors of fund managers to be unusually high.

Recent Basil regulations, such as the liquidity ratio, incentivize banks to borrow "long term." In the context of our model, this regulation can be interpreted as requiring the bank to borrow κ using a 2-period loan (a loan that pays off at date 2) and immediately invest the proceeds in the investment technology that pays off in period 2. If the bank borrowed κ long term and $N - \kappa$ via demand deposits, then allocation $(\widetilde{c}_1, \widetilde{c}_2)$ can be replicated for all $n < N$, eliminating the panic equilibrium. In state $n = N$, however, depositors will only receive

[14] See Kacperczyk and Schnabl (2010).
[15] See: https:// www.sec.gov/ News/ PressRelease/ Detail/ PressRelease/ 1370542347679

$(N-\kappa)y/N < y$ since κ is invested in a technology that pays off in date 2. If the probability of a sunspot, θ, is sufficiently large, such a regulation will generate higher welfare than the efficient allocation (since the later is subject to a bank panic).

Based on their examination of the historical record of financial crises, Calomiris and Gorton (1991) conclude that institutional factors, such as branch bank laws and bank cooperation arrangements, play an important role in determining both the frequency and magnitude of bank panics. Williamson (1989) studies the institutional differences between the banking systems in Canada and the United States during the U.S. National Banking era. The upshot of the historical evidence is that banking systems with fewer but larger banks appear more resilient to panics. Note that "largeness" here should be taken to include the propensity for banks to engage in branch banking activities and to cooperate with other banks in an interbank market. In our model, the parameter indexing the size of a bank or banking system is N. Our model predicts that larger banks (or a more interconnected banking system) are better because larger and more diversified banks are less likely to experience extremely high redemption states, thereby lowering the expected cost associated with extra haircuts in high redemptions states needed to dissuade panic behavior.

We remarked earlier that we viewed the assumption of scale economies as self-evident. This is, of course debatable. Nevertheless, what it is less self-evident why investments with this property are funded with short-term debt. In what follows, we sketch out an extension of our model which suggests that suppliers of financial products are compelled to provide what the market demands. If this is the case, then bank panics might be an inevitable by-product of a competitive financial sector.

Consider an economy with two investment technologies $\{(\kappa_L, R_L), (\kappa_H, R_H)\}$ where $\kappa_L = 0 < \kappa_H$ and $1 < R_L < R_H$. Think of (κ_L, R_L) as representing cash or bonds and (κ_H, R_H) as representing business loans. For a given (κ, R), we can derive the optimal risk-sharing arrangement as described above. Appealing to the results of our analysis above, the cash portfolio is panic free. The business portfolio, on the other hand, may be subject to panics. Let $0 \leq \theta \leq 1$ denote the probability of a panic on the business portfolio financed with demandable debt. Then there is a trade-off for investors to consider: they can invest in the low-return, but panic-free, "narrow bank." Or they can invest in the "shadow bank" that offers a slightly higher rate of return, but is subject to the occasional panic event. A given population N of depositors *with different risk tolerances* is likely to divide itself between these these two banking structures. In this way, financiers might always be able to find a willing clientele for a liquid debt instrument that yields a few basis points more than a liquid and panic-free alternative.

Thus, to the extent that central bank and treasury actions influence real yields on money and bonds, this extended model also suggests there might be some support for the notion that low-interest rate policies (R_L) can induce portfolio substitution by yield-oriented investors into higher-return (R_H) panic-prone debt instruments (Stein, 2013). In this manner, a central bank policy of paying interest on reserves can promote financial stability, since the marginal risk-tolerant investor in our extended model is motivated to substitute into relatively safe (and panic free) central bank liabilities when the interest paid on reserves is increased.

References

1. Beccali, Elena, Mario Anolli and Giuliana Borello (2015). "Are European Banks Too Big? Evidence on Economies of Scale," *Journal of Banking & Finance*, 58(C): 232–246.

2. Bryant, John (1980). "A Model of Reserves, Bank Runs, and Deposit Insurance," *Journal of Banking Finance*, 43: 749–761.

3. Calomiris, Charles and Gary Gorton (1991). "The Origins of Banking Panics: Models, Facts, and Bank Regulation," in *Financial Markets and Financial Crises*, edited by Glenn Hubbard, University of Chicago Press, 109–174.

4. Capiro, Gerard, Jr. and Daniela Klingebiel (1997). "Bank Insolvency: Bad Luck, Bad Policy, or Bad Banking?" in M. Bruno and B. Pleskovic, eds., *Annual World Bank Conference on Development Economics.* Washington DC: World Bank, 79–104.

5. Cooper, Russell and Ross, Tom (1998). "Bank Runs: Liquidity Costs and Investment Distortions," *Journal of Monetary Economics* 41, 27–38.

6. Diamond, Douglas and Philip Dybvig (1983). "Deposit Insurance and Liquidity," *Journal of Political Economy*, 91(June): 401–419.

7. Ennis, Huberto M. and Todd Keister (2010). "On the Fundamental Reasons for Bank Fragility," FRB Richmond *Economic Quarterly* 96, First Quarter: 33-58.

8. Gorton, Gary (1988). "Banking Panics and Business Cycles," *Oxford Economic Papers*, 40: 751-781.

9. Green, Edward and Ping Lin (2000). "Diamond and Dybvig's Classic Theory of Financial Intermediation: What's Missing?" Federal Reserve Bank of Minneapolis *Quarterly Review*, 24 (Winter): 2–13.

10. Green, Edward and Ping Lin (2003). "Implementing Efficient Allocations in a Model of Financial Intermediation," *Journal of Economic Theory*, 109(1); 1–23.

11. He, Jie, Qian, Jun and Philip Strahan (2015). "Does the Market Understand Rating Shopping? Predicting MBS Losses with Initial Yields," *Review of Financial Studies*, 29(2): 457–485.

12. Hughes, Joseph P. and Loretta J. Mester (2013). "Who Said Large Banks Don't Experience Scale Economies? Evidence from a Risk-Return-Driven Cost Function," Federal Reserve Bank of Philadelphia Working Paper No. 13-13/R (forthcoming in the *Journal of Financial Intermediation*).

13. Kacperczyk, Marcin and Philipp Schnabl (2010). "When Safe Proved Risky: Commercial Paper during the Financial Crisis 2007–2009," *Journal of Economic Perspectives*, 24(1): 29–50.

14. Peck, James and Karl Shell (2003). "Equilibrium Bank Runs," *Journal of Political Economy*, 111(1): 103–123.

15. Stein, Jeremy (2013). "Yield-Oriented Investors and the Monetary Transmission Mechanism," remarks presented at "Banking, Liquidity and Monetary Policy," a Symposium Sponsored by the Center for Financial Studies in Honor of Raghuram Rajan.

16. Wheelock, David and Paul W. Wilson (2016). "The Evolution of Scale Economies in U.S. Banking," Federal Reserve Bank of St. Louis Working Paper 2015-021.

17. Williamson, Stephen (1989). "Bank Failures, Financial Restrictions, and Aggregate Fluctuations: Canada and the United States, 1870–1913," Federal Reserve Bank of Minneapolis *Quarterly Review*, Summer: 20–40.