Bank runs without sequential service

Bank runs without sequential service*

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

Ed Nosal
Federal Reserve Bank of Atlanta

July 2018

Abstract

Banking models in the tradition of Diamond and Dybvig (1983) rely on sequential service to explain belief driven runs. But the run-like phenomena witnessed during the financial crisis of 2007–08 occurred in the wholesale shadow banking sector where sequential service is largely absent. This suggests that something other than sequential service is needed to help explain runs. We show that in the absence of sequential service runs can easily occur whenever bank-funded investments are subject to increasing returns to scale consistent with available evidence. Our framework is used to understand and evaluate recent banking and money market regulations.

*We thank conference participants at the 2nd Annual Missouri Macro Workshop, the 2017 Summer Workshop on Money, Banking, Payments, and Finance, at the Bank of Canada, the 2017 Canadian Macro Study Group in Ottawa, as well as seminar participants at the Federal Reserve Banks of Atlanta, Chicago, Cleveland and St. Louis, the National University of Singapore, Arizona State University, University of Hawaii and Simon Fraser University. We owe a special thanks to Todd Keister whose comments on a prior draft led to a number of substantial improvements. The views expressed here are our own and should not be attributed to the Federal Reserve Banks of Atlanta and St. Louis, or the Federal Reserve System. JEL Codes: G01, G21, G28

1 Introduction

The hallmark of financial intermediation is 'borrow short and lend long.' The quintessential example of this is a conventional retail bank that uses demandable debt to finance non-marketable, but otherwise safe, longer-term, higher-return investments. This sort of liquidity mismatch is often identified as one reason why financial intermediaries are subject to belief driven runs. Economic models of bank runs, however, need more than just a liquidity mismatch to explain runs. Diamond and Dybvig (1983) and the large literature that follows appeal to sequential service—the practice of serving depositors on a first-come, first-serve basis and conspicuous in most retail settings—as the ingredient needed to generate bank runs.[1]

Since the 2007-08 financial crisis is widely understood to be a bank run at the wholesale level (Bernanke 2009 and Gorton 2009), it seems natural to interpret the crisis through the lens of the Diamond and Dybvig (1983) model. Such an approach, however, may be misguided because sequential service—an essential ingredient for Diamond and Dybvig—is absent in wholesale banking. To remedy this situation we replace sequential service and consider instead increasing returns to scale in the financial intermediary's investment technology. We show that when the Diamond and Dybvig environment is modified in this manner, bank run equilibria exist in large regions of the parameter space.

The notion that most investments entail a fixed cost seems to us eminently plausible. It has, in fact, been documented for the retail banking sector and it appears to be relevant for wholesale banking. At the retail level, Mester (2008), Wheelock and Wilson (2017), and Corbae and D'Erasmo (2018) provide empirical support for increasing returns to scale in the banking industry. Our analysis, along with the empirical evidence, suggests that factors other than sequential service can lead to retail bank instability. We believe this result is important because there is a pervasive view in the Diamond and Dybvig (1983) literature that absent sequential service, banks would otherwise be stable. We show that this is not the case. Regarding wholesale financial intermediation, given the way that institutions such as shadow banks fund themselves—by rolling over short-term debt—increasing returns to scale represents a good approximation of their investment technology, which we now

[1] Ennis and Keister (2010) provide a comprehensive survey of the literature.

explain.

A leading example of a pre-financial crisis wholesale shadow bank is a dealer bank that funds its assets by borrowing on the overnight repo market. A repo lender provides cash to a dealer bank and receives asset collateral in return. If the tenor of the repo loan is overnight, then the next day the lender returns the asset collateral in exchange for an amount of cash that equals the principle and interest for the overnight loan. These overnight repo loans were typically rolled over on an ongoing basis. This means that at the beginning of the day the lender returns the collateral to the borrower and receives principle and interest—from the previous day's loan—and later on that day provides a cash loan to the borrower in exchange for the asset collateral and so on. This repo borrowing and lending arrangement resembles a demand deposit: rolling over a repo loan looks like the lender keeps its cash "deposited" at the dealer bank and not rolling it over looks like the lender withdraws its deposit on demand from the dealer bank. But, unlike a demand deposit, if the repo lender chooses to withdraw its funds from the repo arrangement—that is, chooses not to re-lend to the dealer bank—the lender or borrower does not face anything that resembles a sequential service constraint.

Repo lenders view themselves as facing an increasing returns to scale investment technology. This view is both consistent with and embodied in the models of Cole and Kehoe (2000) and Gertler and Kiyotaki (2015). To see this suppose a shadow bank funds a large portfolio of non-government securities via repo and that, at least initially, lenders provide sufficient repo loans to fund the portfolio which they roll over. Going forward, the bank is able to fund its securities and lenders receive an agreed upon rate of return. Now imagine that, for some unexpected reason, a large fraction of lenders do not roll over their repo loans and the bank is unable to find new sources of funds. The bank can use the cash received from the small fraction of lenders who did roll over to fund a small part of its portfolio but will have to sell a large fraction of its portfolio in order to fund the rest. If the securities in its portfolio are illiquid, then the sale of a large block of securities will depress the its price. The next day the shadow bank will be unable to reverse the repo transaction with its lenders—since it does not have the resources to do so—implying that the repo lender will be stuck with collateral that is now trading at a depressed price. If the lender is compelled to sell the collateral, the gross return will be low and the net return is likely to be negative. In this way, from the lender's perspective, the shadow bank's investment

technology appears subject to increasing returns to scale. Specifically, if the bank receives large amounts of funding, returns will be high; if it receives small amounts of funding, returns will be low.

We examine the potential fragility of banking structures without sequential service in the most direct and transparent manner possible. In particular, following Green and Lin (2003) and Peck and Shell (2003) we take a mechanism design approach to investigate the question of optimal liquidity insurance when there is aggregate liquidity risk and liquidity preference is private information. We depart from this earlier work by replacing sequential service with a non-convexity in the investment technology. We find that bank runs can easily emerge in the modified environment even under optimal contractual arrangements. Importantly, our optimal contracting approach provides additional support for the outcomes decribed in Gertler and Kiyotaki (2015). Gertler and Kiyotaki (2015) model rollover financing with its implications for asset pricing when debt is not rolled over (similar to the narrative that underlies our increasing returns to scale assumption). But they also assume sub-optimal (simple) deposit contracts. Their simple deposit contract plays a critical role in generating a bank run because it leads to insolvency in some states of the world. An alternative contractual arrangement, one that leaves some positive level of bank equity in every state of the world, may be a preferred arrangement but that eliminates the possibility of runs. Since we take a mechanism design approach our deposit contracts are optimal. This is important because it suggests that bank runs are not necessarily the by-product of ill-designed contractual arrangements.[2]

Our model also provides some insights on recent financial market policy proposals and regulations. For example, our analysis suggests that pricing banks' assets at market prices is not sufficient to prevent runs. We show, however, that market pricing of assets along with imposing minimum levels on bank capital can eliminate runs. These observations are both interesting and relevant in light of recent money market regulations require certain

[2]Moreover, our mechanism design approach does not view a preference for stability as axiomatic. We feel that, in practice, policy makers share this view. For example, there exist contractual structures or regulations, such as financial autarky, that can eliminate bank runs. We conjecture that policy makers prefer some financial instability to financial autarky. We show that there are circumstances where policy makers tolerate financial instability when there exist alternatives—not as extreme as financial autarky—that can eliminate runs.

types of money funds price their assets at market values (NAV pricing) *and* impose "gates" and "fees" on withdrawals. Our framework can be used to interpret other policy choices, such as some aspects of recent Basel III banking regulations.

The intuition that underlies fragility in our environment is straightforward. Suppose that investors/depositors believe that a mass redemption event is likely. Investors know that this means that investments will not be funded to scale, so a low return is likely. In this case, depositors with no pressing liquidity needs have an incentive to misrepresent themselves to the bank and withdraw funds early.[3] In Diamond and Dybvig (1983) depositors run for basically the same reason: they want to avoid a low (negative) return on their investment. But because the rate of return available on the bank's underlying investments is unaffected by scale, the mechanism that initiates the run in Diamond and Dybvig (1983) is different from ours. In Diamond and Dybvig (1983) sequential service implies that the amount of resources left over for latecomers after a mass withdrawal will be very small. Hence, depositors have an incentive to run, even if they do not have a liquidity need, in hopes being at or near the front of the service queue so that they get more resources and, as a result, a better return.

The paper is organized as follows. Section 2 describes the economic environment. In Section 3, we characterize the set of efficient incentive-compatible allocations for economies subject to private information and scale economies in investment opportunities. We establish the existence of run equilibria in Section 4. Section 5 considers a number of applications and extensions. For example, we show that our model provides some support for the notion that low real rates of return on safe asset classes can lead to financial instability through a reach-for-yield behavior. As well, we highlight some policy insights implied by our model and we compare our model and results to others in the literature. We conclude in Section 6.

[3] In terms the shadow bank-repo example, a lender with no pressing liquidity needs will choose not to roll over the repo loan and keep the cash.

2 The model

Our model setting is similar to Green and Lin (2003) and Peck and Shell (2003), both of which take a mechanism design approach to Diamond and Dybvig (1983). The economy has three dates, $t = 0, 1, 2$, and a finite number $N \geq 3$ of *ex ante* (date 0) identical individuals. Each individual receives a preference shock between date $t = 0$ and $t = 1$ that determines type: *impatient* or *patient*. Let $0 < \pi < 1$ denote the probability that an individual is impatient. Let π_n denote the probability that $0 \leq n \leq N$ individuals are impatient. We assume that individual types are *i.i.d.* so that $\pi_n = \binom{N}{n}\pi^n(1-\pi)^{N-n}$. The distribution of types has full support, $0 < \pi_n < 1$ for all n.

Impatient individuals want to consume at date 1 only. Patient individuals 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 individual at date t. Date 0 preferences are given by

$$U(c_1, c_2) = \pi u(c_1) + (1-\pi)u(c_1 + c_2), \tag{1}$$

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

Each individual is endowed with y units of date 1 output. There exists a technology that transforms k units of date 1 output into $F_\kappa(k)$ units of date 2 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.[4] When the minimum scale κ is not met, the rate of return reflects the cost of intermediated storage, indexed by the parameter $1 - r$. Technology (2) generalizes the standard specification used in the literature which assumes $\kappa = 0$ and implies $F_0(k) = Rk$ for all $k > 0$.

There are two benefits associated with cooperation in this economy. First, there are the usual gains associated with sharing risk. Second, and absent

[4] One can easily generalize the analysis to permit multiple threshold levels with associated rates of return. We assume a single threshold level since this is the simplest way to show how our mechanism works.

from the standard model, minimum scale is more easily attained when resources are pooled.

For convenience we adopt the same labels for agents and mechanisms as Diamond and Dybvig (1983) and the literature that follows. We refer to a risk-sharing arrangement that pools resources and exploits scale economies as a *bank.*[5] Individuals who deposit resources with the bank are called *depositors.* A bank can be viewed as a resource-allocation mechanism that pools the resources of the N depositors before they learn their types. In exchange for deposits, the bank issues state and time-contingent deposit liabilities redeemable in output. Because liquidity preference is private information the optimal risk-sharing arrangement includes options to withdraw funds on demand. It is in this sense that the optimal contract resembles conventional demand deposit liabilities (Bryant 1980).

We adopt the conventional island metaphor to describe the structure of communications. In particular, there is a center island and a set of spatially separated islands. Individuals located on the center island can talk to one other. Individuals located on different spatially separated islands are effectively incommunicado.

The timing of events is as follows. At date 0, all N individuals are at the center island and decide whether or not to participate in a risk-sharing arrangement. Participation entails depositing endowment y at the bank and agreeing to the terms of a contract governing the returns on future redemptions.[6] The bank permanently resides at the center island. In between dates 0 and 1, individuals leave the center island and travel to N spatially separated locations. Upon arrival, individuals learn their type: patient or impatient. We assume that depositors can only return to the center island once—either in date 1 or in date 2. This captures the idea that depositors communicate with their bank only when they want to make a withdrawal and that remaining in constant contact with their bank is too costly.[7] When depositors

[5] A "bank" can be a financial intermediary/dealer that funds its assets by overnight repo and "depositors" are investors that provide funding in exchange for repo collateral.

[6] Individuals that choose not to participate consume y at date 1 if $y < \kappa$; if $y > \kappa$, a impatient individual consumes y at date 1 and a patient agent consumers $R\kappa$ in period 2.

[7] If communication was costless between the N individuals or if the individuals could visit the center island at both dates 1 and 2, then they would be able to trade directly with each other, rendering the bank redundant. In appendix 2 we formally describe the communications frictions we impose and compare them to those of the standard models

return to center island (to make a withdrawal from the bank) they arrive simultaneously, not sequentially.

The communications structure implies that date 1 consumption payments specified in the bank contract need only be conditioned on the number of depositors m who visit the bank at date 1, where $m \in \{0, 1, ..., N\}$. In particular, if m depositors visit the bank at date 1, then each depositor receives $c_1(m)$ units of date 1 consumption. Depositors who visit the bank at date 2 each receive $c_2(m) = F_\kappa[Ny - mc_1(m)]/(N - m)$ units of date 2 consumption. Hence, the bank offers depositors a contract in the form of a promised allocation $(\mathbf{c}_1, \mathbf{c}_2)$, where $\mathbf{c}_1 = [c_1(1), \ldots, c_1(N)]$ and $\mathbf{c}_2 = [c_2(0), c_2(1), \ldots, c_2(N-1)]$.

The allocation $(\mathbf{c}_1, \mathbf{c}_2)$ is feasible by construction. However, because liquidity preferences are private information, depositors may want to misrepresent themselves to the bank. To ensure that $(\mathbf{c}_1, \mathbf{c}_2)$ promotes efficient resource allocation, the allocation should be structured in a manner that gives depositors an incentive to represent their preferences truthfully. We restrict attention to economies where it is socially optimal for impatient depositors to consume at date 1 and for patient depositors to consume at date 2. In this case, incentive-compatibility boils down to ensuring that depositors arrive at the bank at a date that corresponds to their type. We now describe the strategic interaction among depositors which we model as a *withdrawal game* that is played after individuals learn their types.

Suppose that all N individuals deposit their endowments with the bank at date 0.[8] In between dates 0 and 1, depositors learn their types and play the following withdrawal simple game: each depositor $j \in \{1, 2, ..., N\}$ simultaneously 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**.** In particular, depositor j does not know the number of impatient depositors n in the economy. A *strategy profile* $\boldsymbol{t} \equiv \{t_1, t_2, ..., t_N\}$ implies an $m \in \{0, 1, ..., N\}$, the number of depositors that return to the bank at date 1. Since the efficient allocation has impatient depositors consuming at date 1

in the literature. We show that the communications frictions that we impose are less restrictive than those needed in the standard literature.

[8] Cooper and Corbae (2002) study an *ex ante* deposit game with increasing returns to intermediation and examine if this game has multiple equilibria. Since we are interested in the *ex post* withdrawal game, we assume that all N individuals participate in the banking arrangement. Below we show that this assumption is without loss of generality.

and patient depositors at date 2, a *truth-telling strategy* has a strategy profile where impatient depositors travel to the center island (or bank) at date 1 and patient depositors travel at date 2. Note that a truth-telling strategy implies that $m = n$.

A strategy profile $\mathbf{t}$ and its associated m constitutes a Bayes-Nash *equilibrium* of the withdrawal game with 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 the truth-telling strategy is an equilibrium for the withdrawal game.

Since $c_1(m) > 0$, it is always a strictly dominant strategy for impatient depositors to visit the bank at date 1. A patient depositor tells the truth by visiting the bank at date 2; he has an incentive to do so—assuming that all other patient depositors visit at date 2—iff

$$\sum_{n=0}^{N-1} \Pi^n u\left[c_2(n)\right] \geq \sum_{n=0}^{N-1} \Pi^n u\left[c_1(n+1)\right], \tag{3}$$

where Π^n is the conditional probability that there are n impatient individuals given there is at least one patient individual and

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

If a feasible allocation $(\mathbf{c}_1, \mathbf{c}_2)$ satisfies (3), then there exists an equilibrium where all depositors play the truth-telling strategy. However, there *may* exist other equilibrium outcomes in the withdrawal game for the feasible allocation $(\mathbf{c}_1, \mathbf{c}_2)$. In particular, there may exist an equilibrium in which depositors play a *run strategy*. A run 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 $n \leq N$.[9]

[9]There is also the possibility of mixed strategy equilibria in which only a fraction of patient depositors misrepresent themselves. We abstract from these "partial runs" because they are peripheral to the main argument we develop below.

3 Efficient incentive-compatible allocations

In this section we characterize the properties of efficient incentive-compatible allocations. We begin with the standard case where the return to investment is invariant to its scale, $\kappa = 0$. We then study the case in which the return to investment is subject to a scale economy, $\kappa > 0$.

3.1 Linear technology

Here we characterize the unconstrained efficient allocation for the linear technology because a subset of this allocation is relevant for the scale economy. The unconstrained efficient allocation is derived under the assumption that depositor types are known which means that impatient depositors visit the bank at date 1 and patient depositors visit the bank at date 2. This implies that n impatient depositors visit the bank at date 1. The unconstrained efficient allocation is given by an allocation $(\mathbf{c}_1, \mathbf{c}_2) \equiv \{c_1(n), c_2(n)\}_{n=0}^{N}$ that maximizes the expected utility of the representative, *ex ante* identical depositor,[10]

$$\max_{\{c_1(n)\}} \sum_{n=0}^{N} \pi_n \{nu\left[c_1(n)\right] + (N-n)u\left[c_2(n)\right]\} \tag{4}$$

subject to the resource constraints

$$\begin{aligned} nc_1(n) &= Ny - k(n) && (5)\\ Rk(n) &= (N-n)c_2(n) && (6) \end{aligned}$$

which when combined yields

$$nc_1(n) + \frac{(N-n)c_2(n)}{R} = Ny, \tag{7}$$

for all $n \in \{0, 1, ..., N\}$, where $k(n) \equiv Ny - nc_1(n)$ represents the resources that remain to fund capital investment. Let $(\mathbf{c}_1^*, \mathbf{c}_2^*)$ denote the solution to the problem above. It is easy to show that there is a unique solution that satisfies

$$u'[c_1^*(n)] = Ru'[c_2^*(n)] \;\forall\; 0 < n < N, \tag{8}$$

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

$c_1^*(0) = c_2^*(N) = 0$ and the resource constraint (7). Given our CES preference specification, the solution is available in closed-form,

$$c_1^*(n) = \frac{Ny}{n + (N-n)R^{1/\sigma - 1}} \tag{9}$$

$$c_2^*(n) = R^{1/\sigma} c_1^*(n), \tag{10}$$

for all $0 < n < N$ with $(c_1^*(0), c_2^*(0)) = (0, Ry)$ and $(c_1^*(N), c_2^*(N)) = (y, 0)$. Note that for all $n < N$ depositors engage in risk-sharing since $y < c_1^*(n) < c_2^*(n) < Ry$. Moreover, because $\sigma > 1$ and $R > 1$ imply $R^{1/\sigma - 1} < 1$, it follows that both $c_1^*(n)$ and $c_2^*(n)$ are decreasing in n.

We immediately have the following result,

Property 1 $c_2^*(n) > c_1^*(n) > c_1^*(n+1)$ for all $n \in \{1, ...N-1\}$.

One implication of Property 1 is that the short and long-term rates of return on deposits, defined as $c_1^*(n)/y$ and $c_2^*(n)/y$, respectively, are both decreasing in the level of date 1 redemption activity, n. Wallace (1988) interprets $c_1^*(n) > c_1^*(n+1)$ as a partial suspension scheme which, by construction, is efficient here.

Using (6), (9) and (10), the efficient level of the date 1 investment, $k^*(n) \equiv Ny - nc_1^*(n)$, is given by

$$k^*(n) = \left[\frac{(N-n)R^{1/\sigma - 1}}{n + (N-n)R^{1/\sigma - 1}} \right] Ny. \tag{11}$$

Notice that $k^*(n)$ is decreasing n. A higher value of n means a higher aggregate demand for early withdrawals. To accommodate this higher aggregate demand, funding for investment is optimally scaled back. 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—this source of return uncertainty of deposits has nothing directly to do with bank fragility. A bank could mitigate the economic impact of these "fundamental runs" by (somehow) expanding its depositor base, N. This could be one of the driving force behind the observed consolidation trend in banking.

There are two important results associated with allocation $(\mathbf{c}_1^*, \mathbf{c}_2^*)$. First, it follows immediately from Property 1 that it $(\mathbf{c}_1^*, \mathbf{c}_2^*)$ is incentive-compatible.

In particular, since $c_2^*(n) > c_1^*(n+1)$ for all $0 < n < N$, allocation $(\mathbf{c}_1^*, \mathbf{c}_2^*)$ satisfies the incentive compatibility condition (3).

Second, the truth-telling equilibrium that implements $(\mathbf{c}_1^*, \mathbf{c}_2^*)$ in the withdrawal 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 for any conjecture $m > 0$, since $c_2^*(m) > c_1^*(m) > c_1^*(m+1)$, i.e., a patient depositor always receives a higher payoff by postponing his withdrawal to the later date. Since it is a dominant strategy for a patient individual to visit the bank at date 2, the allocation $(\mathbf{c}_1^*, \mathbf{c}_2^*)$ can be uniquely implemented as an equilibrium in dominant strategies. We summarize the linear technology case with the following proposition,

Proposition 1 *[Green and Lin, 2000]. The unconstrained efficient allocation $(\mathbf{c}_1^*, \mathbf{c}_2^*)$ is uniquely implementable as a Bayes-Nash equilibrium of the withdrawal game when depositor types are private information and the investment technology is linear.*

Proposition 1 implies that private information and a liquidity mismatch are not in themselves an obstacle to implementing the unconstrained efficient allocation uniquely.[11] It also implies that bank runs do not exist in our environment when investments are subject to constant returns to scale.

3.2 Scale economies

Let $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ and $\hat{\mathbf{k}}$ denote the unconstrained efficient allocation and the associated unconstrained efficient level of investment, respectively, in a scale economy with minimum scale $\kappa > 0$. Note that when $n = N$, achieving scale is irrelevant since all depositors are impatient. Therefore, we always have that $\hat{k}(N) = k^*(N) = 0$ and $\hat{c}_1(N) = c_1^*(N) = y$. Suppose that the minimum scale κ is such that $k^*(j+1) < \kappa < k^*(j)$.[12] These inequalities imply that if there are at least $N - j$ patient depositors, then the investment

[11] Hence, Diamond and Dybvib (1983) and the literature that follows assume sequential service to break the uniqueness result.

[12] Unless κ is either "too high," i.e., $\kappa > k^*(0)$, or "too low," i.e., $\kappa < k^*(N-1)$, these inequalities will be always be valid for some j.

funds available under allocation $(\mathbf{c}_1^*, \mathbf{c}_2^*)$, k^*, will exceed κ. As a result, the return on investment will be R. We therefore have the following property,

Property 2 If (κ, j) satisfies $k^*(j+1) < \kappa < k^*(j)$, then $\{\hat{c}_1(n), \hat{c}_2(n)\} = \{c_1^*(n), c_2^*(n)\}$ for all $n \in \{0, 1, ..., j, N\}$.

That is, the unconstrained efficient allocation in the scale economy corresponds to the unconstrained efficient allocation in the linear economy for all states j, where $k^*(j) \geq \kappa$.

Qualitatively speaking, the remaining allocations $\{\hat{c}_1(n), \hat{c}_2(n)\}$, $n = j+1, ..., N-1$, will take one of two forms. Since $k^*(n) < \kappa$ for $n = j+1, ..., N-1$, the funding for investment associated with these allocations will be characterized by either $\hat{k}(n) = \kappa$ or $\hat{k}(n) < \kappa$. That is, the consumption allocations will be designed so that investment is either exactly equal to minimum scale κ or falls short of it. Since the investment return associated with the former is $R > 1$ and $r < 1$ for the latter, the two allocations take the following form.

If $\hat{k}(n) = \kappa$, then

$$\hat{c}_1(n) = \frac{Ny - \kappa}{n}, \tag{12}$$

$$\hat{c}_2(n) = \frac{R\kappa}{N-n}, \tag{13}$$

$$\hat{k}(n) = \kappa; \tag{14}$$

if $\hat{k}(n) < \kappa$, then we have

$$\hat{c}_1(n) = \frac{Ny}{n + (N-n)r^{1/\sigma - 1}} \tag{15}$$

$$\hat{c}_2(n) = r^{1/\sigma}\hat{c}_1(n), \tag{16}$$

$$\hat{k}(n) = \left[\frac{(N-n)r^{1/\sigma-1}}{n + (N-n)r^{1/\sigma-1}}\right] Ny. \tag{17}$$

Notice that (15)-(17) replicates (9)-(11), respectively, but where R is replaced by r. Since $\sigma > 1 > r$, (15)-(16) imply that $\hat{c}_2(n) < \hat{c}_1(n) < y$, i.e., impatient depositors receive *higher* payments than patient depositors and both payments are *less* than their initial endowment deposit, y. Intuitively,

there is a trade-off between the two options above: investment $\hat{k}(n) = \kappa$ provides a higher total consumption but at the cost of poorer risk-sharing, while investment $\hat{k}(n) < \kappa$ provides better risk-sharing but at the cost of lower total consumption.

To demonstrate our main point we assume that $\kappa = 2y$ and that model parameters that models parameters satisfy $k^*(N-2) < \kappa < k^*(N-3)$.[13] This parameterization implies that the unconstrained efficient allocation for $n \leq N-3$ is given by $(c_1^*(n), c_2^*(n))$, since $k^*(n) > \kappa = 2y$. If, however, $n \in \{N-2, N-1\}$, then the unconstrained efficient allocation will be given by either (12)-(14) or (15)-(17), since in these cases $k^*(n) < \kappa = 2y$.

Let's first examine the case where there are $n = N-2$ impatient depositors. If the "high-return, high-level investment" option, $\hat{k}^H(N-2) = \kappa = 2y$, is chosen, then (12)-(14) imply that $\hat{c}_1^H(N-2) = y$ and $\hat{c}_2^H(N-2) = Ry$, which incidentally corresponds to the autarkic allocation in the standard model. If instead the "low-return, low-level investment" option, $\hat{k}^L(N-2) < \kappa$, is chosen, then (15)-(17) imply $\hat{c}_2^L(N-2) < \hat{c}_1^L(N-2) < y$. Clearly, the high-return, high-level investment option dominates the low-return, low-level investment option since $\hat{c}_1^H(N-2) > \hat{c}_1^L(N-2)$ and $\hat{c}_2^H(N-2) > \hat{c}_1^L(N-2)$. Therefore, we have the following result,

Property 3 For $\kappa = 2y$ and $n = N-2$, the unconstrained efficient allocation in the scale economy is given by the high-return, high-level investment option, where $\hat{k}^H(N-2) = 2y$, $\hat{c}_1^H(N-2) = y$ and $\hat{c}_2^H(N-2) = Ry$.

Now let's examine the case with $n = N-1$ impatient depositors—or, equivalently, one patient depositor. If the high-return, high-level investment, $\hat{k}^H(N-1) = \kappa = 2y$, is chosen (12)-(14) imply that

$$\hat{c}_1^H(N-1) = \frac{N-2}{N-1}y \tag{18}$$

$$\hat{c}_2^H(N-1) = 2Ry. \tag{19}$$

Since $\hat{c}_1^H(N-1) < y < Ry < \hat{c}_2^H(N-2)$, this investment option comes at the cost of very poor risk-sharing. If the low-return, low-level investment

[13] Qualitatively speaking, this parameterization is without loss of generality. See remark 1 below for further discussion. We adopt this particular parameterization because it allows us to easily characterize the unconstrained efficient allocation.

option, $\hat{k}^L(N-1) < \kappa$ is chosen instead, then (15)-(17) imply that

$$\hat{c}_1^L(N-1) = \frac{N}{N-1+r^{1/\sigma-1}}y, \tag{20}$$

$$\hat{c}_2^L(N-1) = r^{1/\sigma}c_1^L(N-1). \tag{21}$$

Inspecting conditions (20) and (21), leads us to the following result,

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

Lemma 1 tells us that if r is close to unity, then the payouts to patient and impatient depositors are approximately equal to y. Let's assume that $r < 1$ is arbitrarily close to unity. Then, by Lemma 1, the expected utility payoff associated with the low-return, low-level date 1 investment option is approximately equal to $u(y)$. Using (18)-(19), the expected utility associated with the the high-return, high-level date 1 investment option is

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

Since this investment option has poorer risk-sharing properties than the low-return, low-level investment option, we would expect the benefit of the former option to diminish with depositors' appetite for risk. Indeed, we can demonstrate that for preferences with $\sigma \geq 2$, the expected utility associated with the low-return, low-level investment option exceeds that of the high-return, high-level investment option, i.e.,[14]

$$\left(\frac{N-1}{N}\right)u\left(\frac{N-2}{N-1}y\right)+\left(\frac{1}{N}\right)u\left(2Ry\right) < u\left(y\right). \tag{22}$$

Therefore, we have the following,

Property 4 For $\kappa = 2y$, $n = N-1, \sigma \geq 2$ and $r < 1$ sufficiently close to unity, the unconstrained efficient allocation in the scale economy is given by the low-return, low-level investment option, where $\hat{c}_1^L(N-1)$ and $\hat{c}_2^L(N-1)$ are determined by (20) and (21), respectively.

[14] See Appendix 1 for the proof.

Property 4 implies that when $n = N-1$, the bank "breaks the buck" in the sense that for every unit that individuals deposit at the bank, they receive less than a unit payoff at date 1, as well as date 2. The empirical relevance of this observation is discussed in Section 5 below.

Properties 1-4 fully characterize the unconstrained efficient allocation in a scale economy parameterized by $k^*(N-2) < \kappa = 2y < k^*(N-3)$, $r < 1$ sufficiently close to unity, and $\sigma \geq 2$. In particular, the unconstrained efficient allocation, $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$, is given by

$$(\{c_1^*(n), c_2^*(n)\}_{n=0}^{N-3}, \\ \hat{c}_1^H(N-2), \hat{c}_2^H(N-2), \hat{c}_1^L(N-1), \hat{c}_2^L(N-1), c_1^*(N), c_2^*(N))$$

We now show that allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ is incentive-compatible. Impatient depositors do not have an incentive to misrepresent themselves, so they always visit the bank at date 1. Regarding patient depositors, in states all states $n \leq N-2$, we have $\hat{c}_2(n) > \hat{c}_1(n+1)$ (from Properties 1, 2 and 3) and in state $n = N-1$, we have $\hat{c}_2(N-1) < \hat{c}_1(N) \approx y$ (from Property 4). Assuming that all other patient depositors visit the bank at date 2, a patient depositor will visit the bank at date 2 if the unconstrained efficient allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ satisfies (3) or, equivalently, if it satisfies

$$\sum_{n=0}^{N-2} \Pi^n \{u[\hat{c}_2(n)] - \Pi^n u[\hat{c}_1(n+1)]\} \geq \Pi^{N-1} \{u[\hat{c}_1(N-1)] - u[\hat{c}_2(N)]\}. \quad (23)$$

Since $\hat{c}_2(n) > \hat{c}_1(n+1)$ for all $n \leq N-2$, the left side is strictly greater than zero. When $r < 1$ is arbitrarily close to unity, the right side is positive but arbitrarily close to zero. Hence, (23) is satisfied with a strict inequality.[15] Therefore, we have the following result,

Proposition 2 *The unconstrained efficient allocation* $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ *can be implemented as a truth-telling equilibrium of the withdrawal game in the scale economy characterized by* $\kappa = 2y$ *and* $\sigma \geq 2$ *with* $r < 1$ *arbitrarily close to 1.*

[15] Since $\hat{c}_2(n) > \hat{c}_1(n) > \hat{c}_1(n+1)$ when $n \leq N-2$, $r < 1$ need *not* be arbitrarily close to unity to have allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ satisfy incentive-compatibility. The condition that $r < 1$ is arbitrarily close to unity simply guarantees that the incentive-compatibility condition (3) will hold with strict inequality. We discuss this in more detail in remark 2, below.

A couple remarks are in order before we proceed to investigate the possibility of run equilibria.

1. While we assume that $\kappa = 2y$, the qualitative properties of the unconstrained efficient allocation remain valid for an arbitrary κ, as long as κ is not "too big" or "too small." In particular, for any $k^*(j+1) < \kappa < k^*(j)$, the solution to the unconstrained efficient allocation entails either $\hat{k}(i) = \kappa$ or $\hat{k}(i) < \kappa$ for $i < j$, where the allocation associated with the latter is characterized by efficient risk sharing. It is straightforward to show that there exists a $\tilde{j} < j$ such that for all $\tilde{j} \leq i < j$, $\hat{k}(i) = \kappa$ and for all $i < \tilde{j}$, $\hat{k}(i) < \kappa$.

2. Property 4 and Proposition 2 both assume that $r < 1$ is arbitrarily close to 1. In this case, we are able to show that the low-return low-level capital investment option in state $n = N-1$ is strictly preferred to the high-return high-level investment capital option and that allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ is incentive compatible. The proposition can clearly remain valid for even lower values of r. To see this, first note that the low-return, low-level capital investment option in state $n = N-1$ is preferred if

$$\left(\frac{N-1}{N}\right) u\left(\frac{N-2}{N-1}y\right) + \left(\frac{1}{N}\right) u\left(2Ry\right) \leq u\left(\frac{Nyr^{1/\sigma}}{1+(N-1)r^{1/\sigma-1}}\right).$$

   The argument inside the right side of the expression above is approximately equal to y when $r \approx 1$ and is strictly increasing in r. Since the above inequality is strict when $r \approx 1$, there exists a $\tilde{r}_{\min} < 1$ so that the left and right sides are equal. Second, note that reducing r from unity does not affect the left side of the incentive compatibility condition (23), but it increases the right side. Therefore, there exists an $\hat{r}_{\min} < 1$ such that (23) is met with an equality. Therefore, both the above inequality and (23), will be met with strict inequalities for any $r \in (r^*_{\min}, 1)$, where $r^*_{\min} = \min\{\tilde{r}_{\min}, \hat{r}_{\min}\}$.

Proposition 2 tells us that allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ can be implemented as a truth-telling equilibrium. We now examine if the truth-telling equilibrium is the *unique* equilibrium for the withdrawal game.

4 Run equilibria

We now investigate if deposit contact $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ generates outcomes other than the truth-telling equilibrium. In particular, we are interested if there exists a run equilibrium defined as $m = N$ for all $n \in \{0, 1, \ldots, N-1\}$. In other words, a run equilibrium is an outcome where all N individuals visit the bank at date 1 regardless of their true liquidity needs. Our main result is reported in the following proposition,

Proposition 3 *For $\kappa = 2y$, $\sigma \geq 2$ and $r \in (r^*_{\min}, 1)$, the unconstrained efficient allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ admits a run equilibrium.*

To check the validity of Proposition 3, propose an equilibrium strategy profile where all N depositors visit the bank at date 1 and ask whether a patient depositor has an incentive to play the proposed strategy. If a patient depositor plays the proposed equilibrium strategy profile and visits the bank at date 1, he receives a consumption payoff equal to $\hat{c}_1(N) = y$. If, instead, he deviates from the proposed equilibrium strategy profile and visits the bank at date 2, then he receives a consumption payoff equal to $\hat{c}^L_2(N-1) < y$ since all other $N-1$ depositors contact the bank at date 1. Clearly, a patient depositor does not have an incentive to deviate from proposed equilibrium play. As a result, a run equilibrium exists.[16]

Note that a run equilibrium can be generated by *any* incentive compatible allocation that satisfies $c_1(N) > c_2(N-1)$. The force of Proposition 3 is to emphasize that such outcomes are not eliminated when we insist that risk-sharing arrangements are efficient. That is, in our environment fragility is not the consequence of suboptimal contractual design.

4.1 Sunspot equilibria and run-proof allocations

Consider the allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ identified in Proposition 3. Would a bank ever offer and depositors accept this risk-sharing arrangement knowing that it is susceptible to a run? Following Peck and Shell (2003) we show that

[16] Notice that if $r = 1$, then a patient depositor would be indifferent between misrepresenting himself or not. Thus, a run equilibrium remains possible even if $r = 1$, though it seems unlikely to survive any reasonable equilibrium refinement.

$(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ can be used to construct a sunspot equilibrium that supports a run equilibrium with strictly positive probability.

A sunspot is an extrinsic event that occurs with probability $0 < \theta < 1$, where the sunspot is observed after individuals have agreed to a risk-sharing arrangement but before they learn their type. A sunspot equilibrium is characterized by a probability θ and the allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$. The equilibrium strategy profile for the sunspot equilibrium is as follows: when the sunspot is *not* observed, an event that occurs with probability $1 - \theta$, depositors play the truthtelling equilibrium strategies described in Proposition 2; when the sunspot is observed, an event that occurs with probability θ, all depositors play the run equilibrium strategies described in Proposition 3. We now verify that a sunspot equilibrium exists for some values of θ.

Let $V(\kappa, R, \theta)$ denote the expected utility associated with allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ when a sunspot occurs with probability θ assuming that depositors play sunspot equilibrium strategies, i.e.,

$$V(2y, R, \theta) \equiv (1 - \theta) E\left[U(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)\right] + \theta u(y).$$

Clearly, $V(2y, R, \theta)$ is strictly decreasing and continuous in θ for all $\theta \in (0, 1]$, with $V(2y, R, 1) = u(y)$. Thus, in contrast to Peck and Shell (2003), a risk-sharing arrangement will always emerge in our environment since $V(2y, R, \theta) > u(y)$ for all $\theta > 0$.[17]

The risk-sharing arrangement that prevails will depend on the probability of a sunspot. In particular, just as in Peck and Shell (2003), if θ is sufficiently large, then the bank may want to eliminate a panic equilibrium by offering an allocation that is *run-proof*. An allocation is run-proof if patient depositors have no incentive to misrepresent themselves when the sunspot is observed. The most efficient way to render an allocation run-proof is for the bank to invest at least κ units of capital in all states $m < N$.[18] When $\kappa = 2y$, the

[17] In our environment, if a sunspot is observed, then each depositor receives a payoff of $u(y)$; if a sunspot is not observed, then the expected payoff to the representative depositor—who does not yet know his type—is $E\left[U(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)\right] > u(y)$. Hence, as long as $\theta < 1$, a banking arrangement will always emerge in equilibrium. In Peck and Shell (2003), if θ is sufficiently high agents will prefer autarky (no-banking) since in their in environment, which assumes a sequential service constraint, the expected utility associated with playing the panic equilibrium is strictly less than $u(y)$.

[18] In state $m = N$, all depositors contact the bank at date 1 so the bank will return the initial depost, y, to the depositors.

efficient run-proof allocation, $(\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2)$, is given by $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ for all $m \neq N-1$ and

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

By construction, allocation $(\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2)$ is incentive compatible and run-proof. It is incentive compatible because $\bar{c}_2(m) > \bar{c}_1(m+1)$ for all $m \leq N-1$. It is run-proof because each allocation in $(\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2)$ is incentive compatible. Specifically, a patient individual has no incentive to visit the bank at date 1 even if he thinks that the other $N-1$ depositors will visit at date 1 since $\bar{c}_2(N-1) = 2Ry > y = \bar{c}_1(N)$.

Let $Q(\kappa, R)$ denote the expected utility associated with the run-proof allocation $(\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2)$, i.e.,

$$Q(2y, R) \equiv EU[(\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2)].$$

Suppose that $Q(2y, R) > u(y)$. Then there exists a $\theta_0 \in (0, 1)$ such that

$$V(2y, R, \theta_0) = Q(2y, R) \tag{24}$$

since: (i) $V(2y, R, \theta)$ is continuously and strictly decreasing in θ; (ii) $V(2y, R, 0) > Q(2y, R)$; and (iii) $V(2y, R, 1) = u(y)$. Hence for any $\theta < \theta_0$, depositors strictly prefer allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ to $(\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2)$, which leaves them exposed to a bank run. If, however, $\theta > \theta_0$, then depositors would prefer the run-proof allocation $(\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2)$ to the run-prone allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$.

Interestingly, it need not be the case that $Q(2y, R) > u(y)$. To see this, notice that for each $n \leq N-2$

$$\left(\frac{n}{N}\right) u[\bar{c}_1(n)] + \left(\frac{N-n}{N}\right) u[\bar{c}_2(n)] > u(y)$$

and that for $n = N-1$ and $\sigma \geq 2$

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

The latter inequality, which is identical to (22), implies that the low-return, low level capital investment option is optimal in state $n = N-1$. Hence, if π_{N-1} is relatively large compared to the other π_j's, then it is possible that $Q(2y, R) < u(y)$, which implies that autarky is preferred to the run-proof

allocation $(\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2)$.[19] If $Q(2y, R) < u(y)$, then the equilibrium outcome for the economy is characterized by allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ since $V(2y, R, \theta) > u(y) > Q(2y, R)$. And, of course, the allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ carries with it the risk of bank runs occurring with probability θ. We summarize the results in the section in the following proposition,

Proposition 4 *If either $Q(2y, R) < u(y)$ or $Q(2y, R) > u(y)$ and $\theta < \theta_0$, then depositors prefer the run-prone allocation $(\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2)$ to the run-proof allocation $(\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2)$; otherwise, depositors prefer the run-proof allocation $(\bar{\mathbf{c}}_1, \bar{\mathbf{c}}_2)$.*

Financial instability arises rather naturally in our environment. The scale economy assumption implies that an individual will always want to run on the bank if he thinks that all other depositors are running. Financial instability will be an equilibrium phenomenon as along as either the probability of a sunspot is not too big or if the expected utility associated with a run-proof allocation is low, i.e., lower than $u(y)$.

5 Discussion

5.1 Recent financial stability regulations

Our model suggests that organizations that fund themselves using very short term borrowing, bank credit lines or commercial paper seem particularly vulnerable to runs. If funding in this form is suddenly pulled in sufficient volume, these organizations could see the value of their long-term operations/investments decline significantly. This, in turn, can reinforce a bleak outlook on the backing of their remaining debt. Something along these lines seems to have occurred on September 16, 2008, when the Reserve Primary Fund "broke the buck." News of this event triggered a large wave of redemptions in the money market sector, especially from funds invested in commercial paper. The wave of redemptions ceased only after the U.S. government

[19] We have assumed, for simplicity, that π_n has a binomial distribution. But our basic results hold for any distribution that has full support. For an arbitrary distribution with full support it is possible to make π_{N-1} relatively large (or small) compared to the other probabilities so that it is possible to you $Q(2y, R) < u(y)$.

announced it would insure deposits in money market funds, essentially rendering them panic-free.[20]

Even though at that time prime investment funds allowed their depositors to withdraw their funds on demand with impunity at a fixed par exchange rate, our model suggests that if these funds were priced using a net-asset-valuation (NAV) method, there might still have been a run. In our model, promised rates of return are made contingent on market conditions, i.e., aggregate redemption demand, and this can be interpreted as a form of NAV pricing of liabilities. Although our efficient risk-sharing arrangement is permitted to break the buck in very heavy redemption states, our flexible NAV-like pricing structure does not in itself eliminate bank runs, even though it improves risk-sharing.

On July 23, 2014, the Securities Exchange Commission announced money market reforms that included the requirement of a floating NAV for institutional money market funds, as well as the use of liquidity fees and redemption gates to be administered in periods of stress or heavy redemption.[21] While our model suggests that NAV pricing of demandable liabilities by itself is not sufficient to prevent runs, the use of liquidity fees and redemption gates is consistent with eliminating panics in our model economy. For example, the difference in consumption levels between run-prone and run-free allocations, $\hat{c}_1(N-1) - \bar{c}_1(N-1) > 0$ described in section 4.1, can be interpreted as a liquidity fee that depositors pay to obtain funds when redemption activity is judged by the directors of a market fund to be unusually high. This liquidity free prevents the bank-panic equilibrium.

Other post-financial crisis regulations also take aim at reducing banks' reliance on short-term borrowing. For example, recent Basel *liquidity ratio* regulations are designed to incentivize banks to borrow longer term. The liquidity ratio requires that banks be able to withstand a significant liquidity outflow for a period of 30 days. A bank is better able to survive such a

[20] See Kacperczyk and Schnabl (2010). Recall that run-proof banking is not necessarily optimal in our framework. It is possible that the Reserve Fund was structured optimally relative to a given prior over possible redemption events. When the high-redemption state was realized, the posterior may have changed in a way that rendered a run-proof structure optimal.

[21] A liquidity fee is a payment that the investor incurs to withdraw funds; a gate limits the amount of funds an investor can withdraw. See https:// www.sec.gov/ News/ PressRelease/ Detail/ PressRelease/ 1370542347679

liquidity event if it lends short, which means that it will receive cash during the liquidity event, and borrows long, which means there is a high probability that is will not have to pay off loans during the liquidity event. In the context of our model, this regulation can be interpreted as requiring the bank to have at least κ units of its loans in the form of long-term 2-period debt. This long term debt pays off at date 2. This implies that the bank will always has at least κ invested in the high return project. An implication of this long-term borrowing requirement is that economy will be panic free. However, this sort of long-term borrowing introduces a new and *additional* cost. In particular, in the event that all N depositors withdraw early for fundamental reasons, which is an event that occurs with low probability when N is large, the bank can only distribute $(N-2)y$ units date 1 consumption because $\kappa = 2y$ is tied up in the long-term investment. This implies that $2y$ units will be wasted.

The above example demonstrates that regulation can eliminate financial fragility associated with runs. But if the regulator's only objective is to eliminate runs then its policies can very well be welfare decreasing. For example, if the above long term borrowing requirement is replaced by the requirement that the bank maintain at least κ resources in long term investment if less than N depositors visit the bank at date 1, then welfare unambiguously increases.[22] Or if $V(2y, R, \theta) > Q(2y, R)$, i.e., an allocation that admits a run generates higher expected utility than the run-free allocation, then the regulator can increase welfare by simply "doing nothing" instead of required long-term borrowing, even though doing nothing may result in a run. Recently, some (Dodd-Frank) regulations in the U.S. have been relaxed. Although this may increase the probability of financial stability, one explanation might be that these regulations are, in fact, imposing costs on the economy and relaxing them increases welfare.

5.2 Reach for yield

Our basic environment with a minor modification can help us put some structure on the concept of *reach for yield*. This term is used extensively in the popular press to describe the idea that investors will choose higher

[22] Welfare increases since $2y$ will not be wasted if all N depositors visit the bank at date 1. Nevertheless, the regulator may require that the financial institution may borrow long—or the financial institution may have an incentive to borrow long—because it may be easier to either monitor and/or implement compared to the alternative strategy.

return, riskier investments when safe interest rates are low. Suppose now there are two fundamentally risk-free investments in the economy parameterized by $\{(\kappa_1, R_1), (\kappa_2, R_2)\}$, where $(\kappa_1, R_1) = (0, \delta)$, $(\kappa_2, R_2) = (2y, R)$ and $1 \leq \delta < R$. The former investment is not subject to a scale economy while the latter is. The reader is invited to think of the (κ_1, R_1) investment as a money mutual fund or repo invested in safe short-term government bonds and the (κ_2, R_2) investment as a prime institutional money fund or repo on non-government securities. For convenience we will refer to the former as safe repo and the latter as risky repo.

Let $(\mathbf{c}_1^{1*}, \mathbf{c}_2^{1*})$ denote the unconstrained efficient allocation associated with the safe repo (linear) investment $(\kappa_1, R_1) = (0, \delta)$. Since $\kappa_1 = 0$, this fund is run-free. Let $(\hat{\mathbf{c}}_1^2, \hat{\mathbf{c}}_2^2)$ denote the unconstrained efficient allocation associated with the risky repo (scale) investment, where $(\kappa_2, R_2) = (2y, R)$. Assume that the probability of a sunspot θ is given by $0 < \theta < \theta_0$, where θ_0 is defined in (24). Since $0 < \theta < \theta_0$, allocation $(\hat{\mathbf{c}}_1^2, \hat{\mathbf{c}}_2^2)$ is exposed to runs and is preferred to a run-free allocation $(\bar{\mathbf{c}}_1^2, \bar{\mathbf{c}}_2^2)$.

Let $V(0, \delta, 0)$ represent the expected utility associated with allocation $(\mathbf{c}_1^{1*}, \mathbf{c}_2^{1*})$ and $V(2y, R, \theta)$ denote the expected utility associated with allocation $(\hat{\mathbf{c}}_1^2, \hat{\mathbf{c}}_2^2)$. Clearly, $V(0, \delta, 0)$ is strictly increasing in δ with $V(0, 1, 0) = u(y)$. Given the properties of $V(0, \delta, 0)$ and $V(2y, R, \theta)$, it is evident that for a given $\theta \in (0, \theta_0)$, there exists a $\delta_0 \in (1, R)$ such that

$$V(0, \delta_0, 0) = V(2y, R, \theta). \tag{25}$$

Hence, there exists a risk-return trade off in our model, but where the risk is extrinsic. Because our model assumes that investors have identical preferences, only one of the two funds will emerge in equilibrium; the one that generates the highest expected utility for depositors will be observed.[23]

We are now in a position to talk about reach for yield. To begin, consider an environment in which $\delta_0 < \delta < R$. In this case, investors prefer the safe repo allocation $(\mathbf{c}_1^{1*}, \mathbf{c}_2^{1*})$ because $V(0, \delta, 0) > V(2y, R, \theta)$. Now suppose that the rate of return on the safe repo declines to δ', where $1 < \delta' < \delta_0$. The cause of this decline in the safe real interest rate is immaterial.[24] The decline

[23] Modeling preference heterogeneity with respect to risk-tolerance would produce a model where risk-sharing arrangements with different risk-return characteristics could coexist. We believe that the intuition that follows in our discussion would survive such a generalization.

[24] Perhaps it is induced by central bank policy as an attempt to bolster the economy in

in the safe real interest rate, whatever its source, will induce a portfolio reallocation. In particular, since $V(2y, R, \theta) > V(0, \delta', 0)$, investors in our model will move their resources out of the safe repo investment into the risky repo investment which offers the allocation $(\hat{\mathbf{c}}_1^2, \hat{\mathbf{c}}_2^2)$ even though the latter is subject to runs. Hence, there is a sense in which our model supports the notion that low real interest rates may lead to a reach-for-yield behavior that renders the financial system less stable and more prone to bank runs, (see Stein 2013).

5.3 Scale economies

We are not the first to appeal to a scale economy to explain financial fragility in financial intermediation. Cooper and Corbae (2002) assume increasing returns to scale in the intermediation process itself. In their model, intermediaries monitor investors and monitoring is subject to increasing returns to scale. In particular, the unit cost of intermediation declines in the investor base and the size of the intermediation sector determined in a non-cooperative participation game. Their environment generates multiple equilibria, with one equilibrium characterized by high entry and investment, and the other by low entry and investment.

We could also have modeled a participation game at date 0 where individuals decide whether to deposit their endowment with the bank or to keep it. Notice that this date 0 participation game is conceptually divorced from the phenomenon of depositor misrepresentation in our withdrawal game since individuals do not know their types at date 0.[25] In our model economy, in contrast to Cooper and Corbae (2002), increasing returns to scale in the investment technology does *not* imply an indeterminacy in the date 0 participation. To see this, notice that the efficient allocation that the bank offers at date 0 is a contingent contract: the bank offers an allocation $[\mathbf{c}_1(\eta), \mathbf{c}_2(\eta)]_{\eta=1}^N$, where η represents the number of individuals that deposit their endowment at the bank at date 0. The expected utility associated with any allocation $(c_1(\eta), c_2(\eta))$ is at least equal to $u(y)$ and is (weakly) increas-

the face of an imminent recession. Alternatively, imagine a wave of pessimism in a part of the globe resulting in a flight to the safety of U.S. treasury debt, leading to a decline in real yields.

[25] Cooper and Corbae (2002, pg. 161) explicitly state that the focus of their paper is not on bank runs.

ing in η. For example, the bank can always offer a contract with allocation $(c_1(\eta), c_2(\eta))$, where $c_1(\eta) = y$ and $c_2(\eta) = 0$. This allocation generates a payoff equal to $u(y)$. In general, the bank offers a contract $[\mathbf{c}_1(\eta), \mathbf{c}_2(\eta)]_{\eta=1}^N$ where allocations differ from $c_1(\eta) = y$ and $c_2(\eta) = 0$ for some η whenever $EU\{[\mathbf{c}_1(\eta), \mathbf{c}_2(\eta)]_{\eta=1}^N\} > u(y)$.

Consider a date 0 equilibrium strategy profile where $\hat{n} < N$ individuals deposit their endowment at date 0 and $N - \hat{n}$ do not, where the bank offers contract $[\mathbf{c}_1(\eta), \mathbf{c}_2(\eta)]_{\eta=1}^N$. Individuals who do not deposit at date 0 receive a payoff equal to $u(y)$ if they follow the proposed equilibrium strategy profile. Do individuals have an incentive to follow this prescribed date 0 strategy profile? If the expected utility of allocation $(c_1(\eta), c_2(\eta))$ is strictly increasing in η and exceeds $u(y)$, then an individual who is not supposed to deposit at date 0 can increase his expected utility by depositing at date 0; an agent who is supposed to deposit will reduce his expected utility if he chooses not to deposit at date 0.[26] The unique date 0 equilibrium strategy profile has all N individuals depositing their endowments at the bank. Hence, in our model economy the increasing returns to scale assumption does not introduce a date 0-indeterminacy in the participation game; all individuals will deposit their endowment at date 0. Thus, the scale economy in our environment does not introduce an additional source of indeterminacy.

5.4 Liquidation as a source of increasing returns

While we have modeled a non-convexity at the level of investment, we could have instead modeled increasing returns in the business of banking itself. In fact, there is an alternative interpretation of increasing returns that makes our model comparable to some related literature.[27]

Suppose that the aggregate endowment Ny is invested entirely as capital at date 0. At for date 1, capital can be *liquidated* at a unit rate of return. As long as the level of liquidation does not exceed a threshold, $Ny - \kappa$, any remaining capital yields a rate of return $R > 1$ in date 2. For levels of liquidation that exceed this threshold, any remaining capital yields a rate of

[26] If the contract allocation $(c_1(\eta), c_2(\eta))$ is weakly increasing in η, then a trembling hand refinement will eliminate all potential equilibrium deposit strategies except for the strategy that all N agents deposit their endowment at date 0.

[27] We thank Todd Keister for suggesting the interpretation that follows.

return $r < 1$. Our scale economy can be reinterpreted as costly liquidation, where the unit cost of liquidation rises if undertaken at a sufficiently high level—something that happens in high redemption states (n close to N). Notice that the liquidation costs manifest themselves in terms of *future* returns, i.e., lower futures returns can be interpreted as higher liquidation costs.

Costly liquidation is prominently featured in the models of Cooper and Ross (1998) and Ennis and Keister (2006). In their environments, liquidating capital at date 1 reduces capital by more than one unit. But any remaining capital provides a rate of return equal to $R > 1$. Relative to our model, this specification for costly liquidation has a very different implication for the existence of bank panics.

In our environment, assuming that $\kappa = 2y$, we have shown that if a patient depositor believes that the other $N-1$ depositors will visit the bank at date 1, he is compelled to follow the crowd and also visit at date 1. In doing so, the patient depositor increases his payoff from $\hat{c}_2^L(N-1) < y$ to $\hat{c}_1(N) = y$. Or, put another way, when the bank liquidates an additional amount of capital, $\hat{k}^L(N-1) > 0$, it provides the patient depositor with an unambiguously *higher* rate of return, since $\hat{c}_1(N) > \hat{c}_2^L(N-1)$.

In the models of Cooper and Ross (1998) and Ennis and Keister (2006), if a patient depositor believes that all other depositors will visit the bank at date 1, he may *not* have an incentive to follow the crowd.[28] If the patient depositor does in fact visit the bank at date 1, then an additional amount of capital k is liquidated. But only a fraction $0 < \lambda < 1$ of this liquidated capital, λk, is available for consumption. If the patient investor instead visits the bank at date 2, his payoff will be Rk. Because $R > \lambda$, costly liquidation *reduces* early consumption relative to late consumption; that is, liquidation is a force that works *against* a patient depositor misrepresenting himself.[29] In contrast, if a patient depositor believes that all other depositors will visit the bank at date 1 in our model, then liquidating the capital for that patient depositor—if he decides to visit the bank at date 1—effectively increases early consumption relative to late consumption. Hence, liquidation *reinforces* the

[28] Cooper and Ross (1998) and Ennis and Keister (2006) assume inefficient contracting, a continuum of depositors, and sequential service. The present discussion assumes a finite number of depositors, that the bank contract is efficient, and that depositors do not know their place in the sequential service queue.

[29] Whether the patient depositor chooses to visit the bank at date 1 or date 2 entails the same sort of calculus as in Peck and Shell (2003).

incentive for a patient depositor to misrepresent himself.

5.5 Deposit insurance and other government policies

One interpretation of Diamond and Dybvig (1983) is that they establish the benefits associated with a deposit insurance scheme. There is some question as to whether their deposit insurance scheme violates their implicit sequential service constraint (Wallace, 1988).[30] If the sequential service constraint does not apply to the government, then Diamond and Dybvig (1983) make the point that deposit insurance, along with a standard deposit contract, can uniquely implement the efficient risk-sharing allocation. But the optimal direct mechanism in our model also generates the unique and efficient allocation result when $\kappa = 0$, just like Diamond and Dybvig (1983). But notice that when $\kappa > 0$ our optimal mechanism does not necessarily rule out the existence of bank runs. As is well known, the mechanism design approach is "institution free," so in the context of our model, one is free to interpret that an optimal deposit insurance program is implicitly embedded in the solution. This implies that there is nothing more that a government can do to improve the properties of the allocation we study without bringing in resources from outside the economy under consideration.

6 Conclusion

The recent financial crisis began with runs on a type of financial intermediary, shadow banks, that is only subtly different from standard deposit taking institutions. As it turns out, this subtly—bank's investors face sequential service constraints while shadow bank's investors do not—requires us to reconsider of our basic understanding of fragility of financial intermediaries. In particular, there a very large literature devoted to studying financial fragility, starting with Diamond and Dybvig (1983), that relies on the sequential service of investors/depositors to generate financial fragility. Indeed, we claim that this literature views sequential service as the *sine qua non* for bank runs.

[30] Like our model, Diamond and Dybvig (1983) consider a closed economy which implies that the deposit insurance scheme must be funded through the existing resources in the economy.

But this view is problematic in light of the recent financial crisis where financial institutions that experienced runs did not face anything that resembled a sequential service constraint.

We provide an explanation for financial intermediary fragility that does not rely on sequential service. In doing so, we view our paper as complementing the existing literature by providing an alternative mechanism—embodied in a financial intermediary's funding structure—that is capable of generating instability. Common to all theories of bank runs is the idea that investors run on financial intermediaries because they believe doing so they will result in higher returns compared to not running. The standard mechanism that generates run behavior is sequential service. Sequential service implies that very little will be left over if there is a mass withdrawal: this provides an incentive for individuals that have no immediate liquidity needs to run on the bank if they think everyone else is running. We offer a new mechanism, increasing returns to scale, that has empirical support in the retail sector. Increasing returns implies that future rates of return will be very low—possibly negative—if there is a mass withdrawal: this also provides an incentive for individuals that have no immediate liquidity needs to run on the bank if they think everyone else is running. Given the funding structure of a shadow bank, from the lenders' perspective, increasing returns to scale seems to be a reasonable characterization of the investment technology they face. Finally, our (mechanism design) approach does not have a built in bias for financial stability and can be used to interpret recent policy and regulatory proposals.

7 References

1. Andolfatto, David, Nosal, Ed and Bruno Sultanum (2016). "Preventing Bank Runs," Forthcoming in *Theoretical Economics*.

2. Bernanke, Ben S. (2009). "Opening Remarks: Reflections on a Year of Crisis," *Federal Reserve Bank of Kansas City's Annual Economic Symposium, Jackson Hole, Wyoming, August 21.*

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

4. Cole, Harold and Timothy Kehoe (2000). "Self-fullfilling Debt Crises,"

The Review of Economic Studies, 67: 91-116.

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

6. Cooper, Russell and Dean Corbae (2002). "Financial Collapse: A Lesson from the Great Depression," *Journal of Economic Theory*, 107: 159–190.

7. Corbae, Dean and Pablo D'Erasmo (2018). "Capital Requirements in a Quantitative Model of Banking Industry Dynamics," unpublished manuscript.

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

9. Ennis, Huberto M. and Todd Keister (2006). "Bank Runs and Investment Decisions Revisited," *Journal of Monetary Economics*, 52(2): 217–232.

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

11. Gertler, Mark and Nobuhiro Kiyotaki (2015). "Banking, Liquidity and Bank Runs in an infinite Horizon Economy," *American Economic Review*, 105: 2011-2043.

12. Gorton, Gary (2009). *Slapped by the Invisible Hand: The Panic of 2007*. Oxford University Press.

13. 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.

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

15. 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.

16. Mester, Loretta (2008). "Optimal Industrial Structure in Banking," Chapter 5 in the *Handbook of Financial Intermediation and Banking*, 133–162.

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

18. 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.

19. Wallace, Neil (1988). "Another Attempt to Explain an Illiquid Banking System: The Diamond and Dybvig Model with Sequential Service Taken Seriously," *Quarterly Review* of the Federal Reserve Bank of Minneapolis, 12(4): 3-16.

20. Wheelock, David and Paul Wilson (2017). "The Evolution of Scale Economies in U.S. Banking," Federal Reserve Bank of St. Louis working paper 2015-021C.

8 Appendices

8.1 Appendix 1

Here we show that

$$\left(\frac{N-1}{N}\right) u\left(\frac{N-2}{N-1}y\right) + \left(\frac{1}{N}\right) u\left(2Ry\right) < u\left(y\right). \tag{26}$$

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

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

Since $1-\sigma<0$, (26) becomes

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

or

$$\left(\frac{N-1}{N-2}\right)^{\sigma-1} + \frac{(2R)^{1-\sigma}}{N-1} > \frac{N}{N-1}. \tag{27}$$

Since $\sigma \geq 2$ and $(N-1)/(N-2) > N/(N-1)$, (27) is a valid inequality, which implies that (26) is also valid.

8.2 Appendix 2

Here we describe our communication frictions. We also relate our communication frictions to the standard environment that assumes sequential service, which itself can be interpreted as a communication friction.

In our environment, N depositors leave the center island at date 0 and travel to their own islands. Communication is not possible between the $N+1$ islands. We assume that depositors who return to the center island at dates 1 and 2 arrive simultaneously—this models the notion that there is a no *within period* sequential service constraint. Suppose that depositors were free to travel to the center island at both dates 1 and 2. Then the bank would be redundant. The N individuals could write contracts among themselves

when they are at the center island at date 0 that replicate the unconstrained efficient allocation. Since they all can return to the center island at dates 1 and 2, they are able to execute the contracts at dates 1 and 2. Hence, a bank is *essential* only if individuals can visit the center island at either date 1 or date 2 and not both, which is what we assume.

By design, our communication frictions imply an *absence* of sequential service. All withdrawing depositors are "together" at the bank's location in date 1: therefore, the bank can condition withdrawals on the total number of visiting depositors.

If one is interested in a model with sequential service that generates a bank run, then sequential service can be interpreted as making our communication frictions more severe, as we now demonstrate.

Suppose first that depositors can visit the bank's location at both dates 1 and 2. (We assume this because many models—Green and Lin (2003) and Peck and Shell (2003)—assume that depositors can visit at both dates.) Assume now that individuals are *always* physically isolated from one another after they depart the center island at date 0, meaning they can never simultaneously be at the same island. Hence, after they depart the center island at date 0 individuals can never communicate with one another. When a depositor visits the center island, he can communicate only with the bank and no other agents (and, of course, when he is not at the center island, he is physically isolated from the bank and other agents and cannot communicate with them.) These communication frictions imply that there can only be one visiting depositor at the center island at any instance at either dates 1 or 2; otherwise (some) depositors would not be physically isolated from one another. Hence, these communication frictions imply that depositors must be served sequentially at both dates 1 and 2. Notice here the essentially of the bank is *not* undermined when individuals are able to visit the bank at both dates 1 and 2 because individuals never have an opportunity to communicate with one another after they depart the center island in date 0.

Sequential service requires a tightening of one aspect of communication frictions between individuals compared to our environment—depositors are never in communication with one another after date 0. But, unlike our environment, sequential service communication friction does not restrict depositors to visit that bank only two times—the bank is essential even if depositors can visit at both dates 1 and 2. Hence, compared to Peck and Shell (2003), it

appears as of one aspect of our communication frictions are tightened—after date 0 individuals can never communicate with one another—and another is loosened—individual can visit that center island at both date 1 and 2. If we require that the bank use an optimal mechanism, we now show that Peck and Shell (2003) must restrict center island visits by individual to either date 1 or date 2, just like in our environment.

Peck and Shell (2003) demonstrate the existence of a run equilibrium when there is sequential service, the technology is linear and when depositors visit the bank's location at dates 1 and 2. Andolfatto, Nosal and Sultanum (2017) show that there exists an indirect contracting mechanism for this environment that *uniquely* implements the efficient allocation—meaning that a bank run equilibrium does not exist. This implies that Peck and Shell (2003) use an inefficient mechanism to generate a bank panic when depositors can visit the bank at dates 1 and 2. However, the mechanism that Peck and Shell (2003) use when they assume that depositors can only visit at either date 1 or 2 *is* efficient (and can also generate a bank run). Therefore, if we restrict our attention to the class of efficient mechanisms that can generate bank runs, the relevant comparison of our environment is with the Peck and Shell (2003) environment that has depositors visiting the bank either at date 1 or date 2. In this case, the communication frictions in our environment are strictly less restrictive than in Peck and Shell's (2003) environment that delivers an "efficient" bank panic.