December 2015

Financial Contracting with Enforcement Externalities

Lukasz A. Drozd and Ricardo Serrano-Padial*

Abstract

Contract enforceability in financial markets often depends on the aggregate actions of agents. For example, high default rates in credit markets can delay legal enforcement or reduce the value of collateral, incentivizing even more defaults and potentially affecting credit supply. We develop a theory of credit provision in which enforceability of individual contracts is linked to aggregate behavior. The central element behind this link is enforcement capacity, which is endogenously determined by investments in enforcement infrastructure. Our paper sheds new light on the emergence of credit crunches and the relationship between enforcement infrastructure, economic growth, and political economy distortions.

Keywords: enforcement, credit rationing, costly state verification, state capacity, financial accelerator, credit crunch, global games, heterogeneity

JEL codes: D82, D84, D86, G21, O16, O17, O43.

*Drozd: Federal Reserve Bank of Philadelphia and the Wharton School of the University of Pennsylvania. Serrano-Padial: School of Economics, Drexel University. We thank Jason Donaldson, Patrick Kehoe, Marzena Rostek, Joel Sobel, Marek Weretka, and audiences at Drexel, Indiana, UNC-Chapel Hill, Oxford, University of Pennsylvania, Pompeu Fabra, Philadelphia Fed, Queen Mary, Royal Holloway, St. Louis Fed, Surrey, Wharton, and UW-Madison for their comments. Drozd acknowledges the financial support of the Vanguard Research Fellowship from Rodney L. White Center for Financial Research. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. This paper is available free of charge at www.philadelphiafed.org/research-and-data/publications/working-papers.

1 Introduction

Enforcement of financial contracts often depends on the collective behavior of agents, leading to a feedback loop that can propagate economic shocks and exacerbate the effect of other frictions. The 2007-2010 foreclosure crisis in the U.S. is a stark illustration that contract enforceability is *endogenous* to the state of the economy. During the crisis, a high foreclosure rate made it increasingly difficult for lenders to foreclose in a timely way on delinquent borrowers, delaying the threat of eviction by as much as three years in troubled areas.[1] The implicit transfer associated with foreclosure delays incentivized even more defaults, raising the national foreclosure rate by at least 25% and adversely affecting the mortgage market.[2]

The foreclosure crisis is hardly an isolated episode. In fact, most financial crises test the limits of the enforcement infrastructure and are characterized by a severe contraction of credit.[3] Not surprisingly, the idea of propping up the economy's resolution mechanisms has always been the cornerstone of policy response to financial crises. More generally, there is a well-established link between access to credit and the efficacy of contract enforcement.[4] In light of this evidence, it is important to understand the implications of endogenous enforceability for credit provision,

[1]The national average delay jumped from about 13 months to over 21 months (Calem et al., 2015). In the midst of the crisis, it took almost three years to foreclose in states such as Florida and New York, and the District of Columbia. See "Delaying foreclosure: Borrowers keep homes without paying," CNN Money, Dec. 28, 2011, http://money.cnn.com/2011/12/28/real_estate/foreclosure/.

[2]Using a quantitative model, Herkenhoff and Ohanian (2012) estimate that delays added 25% to the delinquency rate during the crisis. In a related study, Chan et al. (2015) estimate that a delay of nine months was associated with a 40% higher default rate while controlling for a wide array of confounding factors. The work by Mayer et al. (2014) confirms that strategic considerations are an important factor determining households' default decisions. See Cordell et al. (2013) for a thorough discussion of foreclosure delays.

[3]For instance, the volume of nonperforming loans during the Asian crisis reached 14, 27, 32, and 51% of GDP for Korea, Indonesia, Malaysia, and Thailand, respectively (Woo, 2000). Liquidation of these assets took years, and no significant improvement in credit conditions occurred before this was resolved (Stone, 2000; Enoch et al., 1998). In the case of the 1995 Mexican crisis, Krueger and Tornell (1999) argue that delayed resolution was the chief reason responsible for the slow recovery of the Mexican economy.

[4]Empirical studies on the effect of enforcement on credit include Safavian and Sharma (2007) and Jappelli et al. (2005), who use firm data to show a strong correlation between efficiency of enforcement (courts) and access to credit. Using a natural experiment, Iverson (2015) presents direct evidence that Chapter 11 restructuring outcomes are severely affected by delays in busy bankruptcy courts. In a related paper, Ponticelli (forthcoming) shows that court congestion changes the impact on firms' access to credit of financial reforms aimed at strengthening lender protection. For cross-country evidence, see Djankov, Hart, McLiesh and Shleifer (2008).

and elucidate its role in the transmission of shocks and in the development of financial markets.

To this end, we develop a theory of credit supply with endogenous enforceability. Our framework is based on the idea of costly state verification (Townsend, 1979), the workhorse model of financial contracting with enforcement frictions in much of the economics literature. However, unlike in the original framework, enforcement of individual contracts is determined by an economywide enforcement capacity, which must be accumulated ex ante and can become constraining ex post. In our model, agents are aware that when the enforcement capacity is binding, any marginal increase in the aggregate default rate depletes existing capacity and weakens enforceability, strengthening the incentives to default. We refer to this central feature of our model as an *enforcement externality* and study its impact on optimal debt contracting.

Enforcement externalities introduce a coordination problem among borrowers that can have significant effects on the flow of credit in the economy; however, these effects are not well understood. The papers that consider optimal debt contracting (Townsend, 1979; Gale and Hellwig, 1985) as well as quantitative models of financial frictions, such as the financial accelerator model of Bernanke et al. (1999), exhibit exogenous enforceability by assuming that default costs and enforcement are scale invariant. This is also the case in the recent literature studying the political determinants of a nation's enforcement institutions and their effect on economic development (Besley and Persson, 2009, 2010). At the same time, the scant literature that does study enforcement externalities either does not analyze its effect on credit supply, as in the sovereign crisis model of Arellano and Kocherlakota (2014), or focuses on stylized environments featuring homogeneous agents such as the microfinance model of Bond and Rai (2009).[5] Arguably, agent heterogeneity is a prominent feature of models of financial frictions, and it critically shapes the

[5] Arellano and Kocherlakota (2014) study the spillovers of private default to sovereign default in a model with fixed loan size featuring an exogenous constraint on asset liquidation. Bond and Rai (2009) explore lending in a microfinance model of borrower runs. The run is generated by the fact that when more agents choose to default, they exhaust the lender's resources and take away her ability to sustain future borrowing. Coordination problems arising from limited enforcement have also been studied in the crime (Bond and Hagerty, 2010) and tax evasion literatures (Bassetto and Phelan, 2008).

coordination problem faced by agents. Hence, by developing a framework to analyze enforcement externalities under general forms of agent heterogeneity, our paper extends the domain of applicability of existing analysis to a much broader spectrum of economic environments.

Formally, our model involves a principal-multiagent debt contracting problem, similar to the bilateral setup of Gale and Hellwig (1985). There is a population of entrepreneurs with access to risky projects that seek funding from the principal. Prior to issuing credit, the principal invests in enforcement capacity. The principal makes optimal loans to entrepreneurs, who privately observe their project's *heterogeneous* returns after investing the funds and simultaneously decide whether or not to repay the loans. Contracts are enforced after default decisions are made, subject to an enforcement capacity constraint that limits the number of defaulted contracts the principal can enforce. In the absence of enforcement, some agents may have an incentive to default in order to take over a fraction of the project's liquidation value. Hence, the possibility that enforcement capacity becomes binding makes default decisions strategic complements and gives rise to the enforcement externality. Finally, default is socially wasteful, and the principal would like to minimize it by choosing capacity and credit levels.

Methodologically, our model overcomes two challenges. First, strategic complementarities introduce multiple equilibria when the principal's enforcement capacity is observed by the agents, preventing us from deriving the principal's choice of enforcement capacity and credit provision. In response to this issue, we resort to global games methods introduced by Carlsson and van Damme (1993), Morris and Shin (1998), and Frankel, Morris and Pauzner (2003) and select a unique equilibrium by introducing noisy information about capacity. Second, heterogeneity complicates the characterization of the unique equilibrium. To solve this problem, we build on the work of Sakovics and Steiner (2012) on games with symmetric equilibria and generalize their insight to games in which heterogeneity naturally leads to asymmetric equilibria.[6] This

[6]Sakovics and Steiner (2012) allow for heterogeneous preferences but impose preference restrictions to ensure that equilibria are symmetric. In the context of our model, this would imply that either all agents or none

approach allows us to prove our main theoretical result: the characterization of equilibrium default strategies as a function of credit levels and enforcement capacity.

Our analysis reveals that despite heterogeneity in default propensities (induced by heterogeneous project returns), equilibrium exhibits a surprising amount of economic fragility. Specifically, we show that for typical return distributions, a cluster of agents follows the same default strategy, even if their propensity to default is widely heterogeneous. As a result, a wave of socially wasteful default ensues when enforcement capacity falls below a certain critical level, with such level being increasing in the amount of credit provided by the principal.

The clustering of default decisions is generated by a domino effect driven by agents' divergent beliefs about enforcement levels. Agents who are intrinsically less prone to default (i.e., those with high project returns) expect a higher default rate than do more prone-to-default agents, even if their beliefs about capacity are roughly similar. This is because low-propensity types anticipate that whenever they find default optimal, so do higher propensity types. In that case, lower propensity types expect less enforcement. Conversely, high-propensity types expect a lower default rate when they are indifferent between defaulting or not because they assume that lower propensity types must be repaying, thus expecting more enforcement. If these belief differences are big enough, they induce higher and lower propensity types to default at the same time, and a default cluster arises. Belief differences increase with the relative presence of each type in the population. Hence, clustering generally emerges for return distributions that exhibit areas of high concentration of returns, as is the case with the unimodal distributions used in practice, such as the lognormal or the Pareto distribution.

Clustering has important implications for the functioning of credit markets. The first implication is that the discontinuity of the default rate with respect to enforcement capacity makes the economy fragile to shocks that raise insolvency (or reduce enforcement capacity) since they

default. In contrast, equilibrium in our model typically involves some agents repaying and some defaulting.

can trigger a wave of coordinated defaults. The second implication is that, because of the deadweight loss associated with defaulting, reduced enforcement capacity can lead to credit rationing by imposing *hard* borrowing constraints (a.k.a. credit crunch) rather than *soft* interest rate hikes. Specifically, if existing loans default more frequently due to a shock and divert enforcement infrastructure away from the new loans market, the optimal response of the principal is to tighten the supply of credit to new borrowers in order to reduce their intrinsic propensity to default and prevent a default wave. This prediction is relevant for monetary policy, because it implies that lowering interest rates is not necessarily the most effective strategy in this case. Instead, unconventional policies might be more effective. Finally, clustering lowers the effectiveness of capacity buildup in raising credit levels and thus increases the opportunity cost of investments in capacity. Consequently, enforcement externalities can have a strong amplification effect of political economy distortions on economic development by disincentivizing the buildup of enforcement institutions needed to sustain credit markets.

To illustrate these implications, we present two numerical applications of our theory. The first application is in the spirit of the financial accelerator model (Bernanke et al., 1999). In a reasonably parameterized model, we show that a shock that exogenously doubles the default rate on existing loans, as was the case during the 2007-2010 financial crisis in the U.S., may lead to a significant contraction of credit of about 30%. Importantly, we show that if capacity is binding in normal times, credit levels are quite sensitive to default rate fluctuations regardless of the likelihood of these shocks, either because the principal does not have enough capacity to prevent a credit drop when the shock hits (low-probability shock) or because the principal increases capacity in anticipation of the shock, thereby reducing borrowing constraints in normal times (high-probability shock). We refer to this effect as an *enforcement accelerator*.[7]

[7]While the financial accelerator model by Bernanke et al. (1999) also studies the effect of an externality, the nature of their externality is quite different from ours. It arises because the net worth of entrepreneurs is linked to the price of investment/capital goods. In contrast, in our case, the source of the externality is limited capacity to liquidate assets, and it can materialize both before and after the contractual relation. This distinction is

The second application illustrates the impact of enforcement externalities on other frictions. Specifically, we look at the impact of political economy distortions on the efforts of developing economies to build enforcement infrastructure. This application is inspired by the recent workhorse model of political economy and development (Besley and Persson, 2009, 2010, 2011), which is centered on the idea that accumulating enforcement infrastructure or *state capacity* is essential to prop up financial markets and spur economic development. In this context, our findings indicate that enforcement externalities magnify the effect of political economy distortions and make it highly nonlinear. In particular, when externalities are strong, even small distortions on the principal's objective of maximizing welfare can make an economy fall into a *financial development trap*, characterized by little access to credit.

Our analysis provides additional policy recommendations that are worth emphasizing. During periods of financial distress, debt sustainability is a concern because of enforcement externalities. In such a case, it may be desirable to combine government policies aimed at increasing enforcement with interventions targeting financial relief to the least solvent agents (the most prone to default). These policies can be cost-effective in reducing the risk of default clustering by tampering down the externalities. In addition, financial aid targeting the enforcement infrastructure of developing countries may be an effective way of propping up credit markets and compensating for the lack of internal incentives to build this infrastructure when political economy distortions are present.

Beyond the specifics of our framework, our equilibrium characterization technique deals with arbitrary distributions of default propensities, and hence can be used in alternative models of default spillovers—e.g., those generated by endogenous collateral values as in the fire-sales literature (Shleifer and Vishny, 2011) and, in general, to models of strategic complementarities featuring binary actions and heterogeneous preferences.

important, as it addresses the criticism of financial accelerator models raised by Carlstrom et al. (2016).

2 Environment

There are three periods and two types of agents: a benevolent principal and a population of ex ante identical agents (entrepreneurs) of measure one. Agents are risk neutral and need resources to finance their risky investment projects. The principal has deep pockets and can provide funding to agents.

Funding is restricted to loans (debt). Loans are characterized by a tuple $(b, \bar{b})$, where b is the amount that agents take from the principal, and $\bar{b}$ is the amount agents are required to repay to the principal. Agents have their own equity equal to y already invested in the project, implying a total investment of $y + b$. The return on investment is $w - 1$, where w is a random variable that is privately observed by agents (output is $(y + b)w$).

Assumption 1. *1) w has a continuous cdf F with density f and full support in $[0, \infty)$; 2) $\dfrac{F(w)}{wf(w)}$ is increasing and $\lim_{w \downarrow 0} \dfrac{F(w)}{wf(w)} < 1$; 3) $Ew > 1$.*

Property 2) is satisfied by commonly used distributions such as the lognormal and the Pareto distributions and facilitates the interpretation of our results since it leads to the existence of a single default cluster. However, as we emphasize in what follows, our approach to equilibrium characterization applies to any F.[8]

Agents are protected by limited liability and thus may default on what they owe to the principal. When an agent defaults, it is assumed that her project is liquidated. Liquidation diminishes the value of the project to a fraction $0 < \mu < 1$ of its original value (output), but it makes the project transferable to the principal.

The principal captures the full liquidation only when she monitors the project and verifies

[8] We work in our proofs with a discrete distribution of project returns. However, to ease exposition, we lay out our model as a limit of a discrete distribution economy that is an arbitrarily fine approximation of the continuous distribution F. A discrete distribution of types is convenient from a technical point of view. It guarantees that the global game played by agents exhibits dominance regions (i.e., intervals of signals about capacity at which agents have a dominant strategy). The presence of dominance regions, along with i.i.d. signal noise, ensures that a unique strategy profile survives iterated elimination of strictly dominated strategies.

the realization of w. If liquidation is *not* monitored, the agent gets a $0 < \gamma < 1$ fraction of the project's liquidation value and the principal gets the rest. The parameter γ captures in a reduced form the possibility that any defaulting agent eventually loses the project but that the lack of prompt enforcement actions allows the extraction of rents from the project. As will become clear below, a higher γ leads to stronger enforcement externalities by increasing the benefits of unmonitored defaults. The principal accumulates *enforcement capacity* X_o in the first period, which determines the mass of defaulting agents the principal can monitor later on.

To make the model interesting, the following assumption guarantees a finite credit provision in equilibrium. It also implies that the deadweight loss from defaulting is high enough so that loans cannot be fully recouped through liquidation.

Assumption 2. $\int_{w'}^{\infty} w' dF(w) + \mu \int_{0}^{w'} w dF(w) < 1$ *for all* $w' \geq 0$.

Next, we lay out the timing of events and specify actions and payoffs of all agents.

2.1 Timing, constraints, and payoffs

1. Capacity accumulation: At this stage, the principal chooses how to allocate her endowment of resources $R > 0$ between her own consumption and enforcement capacity X_o. The cost of building X_o is $c(X_o)$, where $c(\cdot)$ is convex, continuous, and increasing. Formally, the principal chooses $X_o \geq 0$ and her consumption $g \geq 0$, subject to

$$R = c(X_o) + g. \tag{1}$$

After this decision has been made, an aggregate shock is observed. This shock depletes a random amount $s > 0$ of the principal's capacity X_o, and in a crude way, it captures the idea that the principal has some preexisting loans on her books; when these loans default more frequently, less capacity is available to support new loans. The residual capacity is $X = X_o - s$.

In this *baseline timing*, the shock occurs before the contracting stage. However, we also discuss the *alternative timing* in which the shock arrives after contracts have been signed.

2. Contracting: During the contracting period, each agent applies for a loan $(b, \bar{b})$ with the principal to invest in her project. Since the repayment amount $\bar{b}$ can be mapped to a cutoff level productivity $\bar{w}$ such that $\bar{b} = (y + b)\bar{w}$, from now on we represent loans by the tuple $(b, \bar{w})$.[9]

3. Enforcement: In the last period, project returns are privately observed by agents, and they simultaneously decide whether to default or repay their loans. As already mentioned, if an agent decides to repay her loan, she keeps the project. Her payoff from doing so is

$$w(y + b) - \bar{b} = (y + b)(w - \bar{w}),$$

while the payoff for the principal is $\bar{b} = (y + b)\bar{w}$. If, however, an agent decides to default, the project is liquidated and its value drops to $\mu(y + b)w$. How the liquidation value is split between the principal and the agent depends on whether it is monitored or not. If liquidation is not monitored, the payoff for a defaulting agent is $\gamma\mu(y + b)w$, while the principal gets $(1 - \gamma)\mu(y + b)w$. If liquidation is monitored, the agent gets nothing, and the principal receives $\mu(y + b)w$. Accordingly, the agent's utility function depends on the agent's decision to repay $a \in \{0, 1\}$ ($a = 1$ means repayment) and on monitoring $m \in \{0, 1\}$ ($m = 1$ means monitoring):

$$u(a, w, m) := \begin{cases} (y + b)(w - \bar{w}) & a = 1 \\ \gamma\mu(y + b)w & a = 0 \text{ and } m = 0 \\ 0 & a = 0 \text{ and } m = 1. \end{cases} \tag{2}$$

[9] We do not analyze the optimality of debt contracts but impose them as a restriction on the allocation. While it has been shown that in bilateral contracting with costly state verification debt is indeed optimal (Krasa and Villamil, 2000), it does not apply directly to our multilateral contracting setting.

After the repayment decisions are made, the principal uses her enforcement capacity to monitor agents who did not repay. The default rate ψ is given by

$$\psi := \int_{\{w:a=0\}} dF(w). \tag{3}$$

Given that the principal can monitor at most a measure X of agents, when $\psi > X$, she randomly selects whom to monitor among the pool of defaulting agents.[10] Accordingly, the probability that a liquidated project is monitored is given by

$$P := \min\{X/\psi, 1\}. \tag{4}$$

The constraint on the enforcement probability faced by a defaulting agent is endogenous and depends on the actions that agents take in equilibrium through ψ. This *enforcement capacity constraint* is what gives rise to strategic complementarities in default and to an enforcement externality: As more agents default, the constraint tightens, pushing down expected enforcement levels and making default more attractive.

Having described the timing and payoffs, we now turn to the optimization problems of the agents and the principal. All proofs are relegated to the Appendix.

2.2 The agent's problem

Since agents are homogeneous at the contracting stage, the principal offers the same loan $(b, \bar{w})$ to each agent. As will become clear below, the principal internalizes agents' individual rationality constraint. Hence, the only relevant decision that each agent makes in our model is whether or

[10] We assume that the principal cannot design contracts that profile agents during enforcement based on payoff-irrelevant characteristics, such as the address, names, etc. Such contracts would violate sequential service constraints (Diamond and Dybvig, 1983) as well as nondiscrimination laws, so we do not consider them here. Segmented monitoring (discrimination) is beneficial in our environment and is known to unravel the externality. This has been shown by Carrasco and Salgado (2014).

not to repay the loan after observing her project return. An agent's repayment decision a solves

$$\max_{a\in\{0,1\}} \{P^e u(w,a,1) + (1-P^e)u(w,a,0)\}, \tag{5}$$

where P^e denotes the expected monitoring probability the agent faces upon default.

The above problem implies that agents repay their loans if P^e exceeds a well-defined cutoff value $\theta_{\bar{w}}(w)$. We refer to $\theta_{\bar{w}}(w)$ as the *propensity to default* of an agent of type w.

Lemma 1. *The default decision an agent with contract* $(b, \bar{w})$ *is*[11]

$$a = \begin{cases} 1 & \text{if } \ P^e \geq \theta_{\bar{w}}(w) \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where

$$\theta_{\bar{w}}(w) := 1 - \frac{1}{\mu\gamma}\left(1 - \frac{\bar{w}}{w}\right). \tag{7}$$

Observe that $\theta_{\bar{w}}(w)$ is increasing in w, implying that agents are more prone to default the lower their productivity w is. Hence, heterogeneity of returns translates into heterogeneity of default propensities. Consequently, when all agents expect the same monitoring probability $P^e = P$, the aggregate default rate ψ defined in equation (3) is equal to $F(\hat{w})$, where $\hat{w}$ solves $P = \theta_{\bar{w}}(\hat{w})$. However, when agents receive a noisy signal about X, as in our global game version of the model, P^e differs across agents so understanding how these expectations are formed will be key to solving for equilibrium. Also note that the types whose default decision is driven by their expectations about enforcement levels are those with default propensities between 0 and 1. These are the agents who behave strategically because their incentives to default depend on their beliefs about the behavior of other agents (i.e., about the default rate). Accordingly,

[11] Given the continuity of F, we can assume without loss that an indifferent agent always chooses to repay.

their presence in the population will determine the impact of strategic complementarities on equilibrium default rates and hence the strength of enforcement externalities.

Lemma 2. *The range of agent types with* $\theta_{\bar{w}}(w) \in (0,1)$ *is* $[\bar{w}, \bar{w}/(1-\gamma\mu))$. *Agent types outside this range either never default* ($w \geq \bar{w}/(1-\gamma\mu)$) *or always default* ($w < \bar{w}$).

The lemma implies that the share of strategic agents, given by $F(\bar{w}/(1-\gamma\mu)) - F(\bar{w})$, is a function of three factors: the loan contract ($\bar{w}$), agent heterogeneity (F), and the payoff from not being monitored ($\gamma\mu$). (The proof of the lemma is trivial and therefore omitted.)

2.3 The principal's problem

The principal takes the behavior of agents as given and maximizes agents' expected utility and her own ex ante consumption g weighted by an exogenous preference parameter $\alpha \geq 0$, where we associate $\alpha = 1$ with a benevolent principal who merely recognizes the opportunity cost of forgone public consumption. The case of $\alpha > 1$ captures political economy distortions or other frictions that bias the principal toward consumption and away from accumulation of enforcement capacity.[12]

The principal's optimization problem is sequential. In the first period, the principal allocates ex ante resources R between g and X_o. In the second period, given the realization of the aggregate shock s, the principal chooses contract terms $(b, \bar{w})$. Note that, under the baseline timing, the contract is effectively contingent on X.

The payoff from these two decisions is connected by the preference parameter α. Hence, it could be thought of as a choice problem of two entities that may or may not fully internalize the problem of the other entity. For example, the first stage may describe the choice of a government that builds up an economywide enforcement infrastructure, and the second stage

[12] Conversely, $\alpha < 1$ can be interpreted as a bias in favor of promoting private investment at the expense of alternative uses of public resources.

describes a competitive lending industry that takes the enforcement infrastructure as given. An alternative interpretation is that both decisions are made by the lending industry, but because of some kind of market imperfection, not all the tradeoffs of accumulating capacity are internalized ($\alpha > 1$).

Formally, in the first period the principal chooses X_o and g to solve

$$\max_{X_o,g}[\alpha g + \mathbb{E}\Pi(X_o - s)] \tag{8}$$

subject to $R = c(X_o)+g$ and $g \geq 0$, where $\Pi(X)$ denotes the net expected utility of entrepreneurs conditional on residual capacity $X = X_o - s$. In the second period the principal chooses contract $(b, \bar{w})$ given X to maximize the utility of entrepreneurs, i.e.,

$$\Pi(X) \;:=\; \max_{b,\bar{w},P}\left[\int_{\{w:a=1\}}(y+b)(w-\bar{w})dF + (1-P)\int_{\{w:a=0\}}\gamma\mu(y+b)wdF\right], \tag{9}$$

subject to resource feasibility

$$b \leq \int_{\{w:a=1\}}(y+b)\bar{w}dF + P\int_{\{w:a=0\}}\mu(y+b)wdF + (1-P)\int_{\{w:a=0\}}(1-\gamma)\mu(y+b)wdF, \tag{10}$$

and to the enforcement capacity constraint given by (4).

Observe that in the second period the principal maximizes agents' net expected payoff. That is, even if agents behave opportunistically and default strategically, their utility is fully internalized by the principal. It is the deadweight loss from defaulting that makes default undesirable. In addition, the principal must still break even in expectation, as implied by (10).

Notice that, by construction, there is no need to build enforcement capacity when $\gamma = 0$. In such a case, an agent's payoff under default is unaffected by monitoring, and her repayment behavior is driven solely by $\bar{w}$ as in standard models of costly state verification. It is when γ

is not too low that providing a significant amount of credit requires monitoring to prevent the deadweight loss associated with strategic defaults. This is because, by Lemma 2, the pool of agents who would default strategically in the absence of monitoring is larger at higher credit levels (higher $\bar{w}$) and higher γ. Consequently, in our model building enforcement capacity is tied to the presence of enforcement externalities.[13]

Under the alternative timing of events, the problem of the principal needs to be modified to reflect the fact that the choice of contracts occurs before s is observed. Since it would be straightforward to set up such a problem, we omit its explicit formulation.

3 Analysis

We first analyze the last period and focus on the equilibrium of the *enforcement game* between the principal and the agents. Next, we derive the relationship between credit level and enforcement capacity. Finally, we look at the optimal choice of enforcement capacity and loan contracts as a function of capacity shocks (s) and political economy distortions (α) in Section 4.

3.1 Equilibrium of the enforcement game

At this stage, the values of g, X, $\bar{w}$, and b are given and agents simultaneously choose their action a, while the principal allocates X uniformly to monitor those who default.

It is not difficult to see that under common knowledge of X, the presence of enforcement capacity constraint (4) can lead to a multiplicity of equilibria for a wide range of X. Intuitively, if agents anticipate a default rate ψ higher than X they expect (4) to bind, which can lead to a monitoring probability P low enough to incentivize default by a mass ψ of agents. Similarly, a

[13] It can be formally shown that for any $\gamma > 0$, we can find a loan amount sufficiently close to zero that can be sustained with $X = 0$. The reason is that the mass of strategic agents shrinks to zero as $\bar{w}$ goes to zero. If we modify the model so that defaulting agents keep the full *unliquidated* value of their projects when they are not monitored, then any positive loan amount would require $X > 0$.

belief in a default rate $\psi < X$ can be self-fulfilling given that agents expect to be monitored with probability one. Except at very low and very high levels of capacity there typically exist three equilibria: an efficient equilibrium with no strategic defaults ($\psi = F(\bar{w})$) and two inefficient equilibria with higher default rates (see the Appendix for details).

Multiplicity introduces indeterminacy in the principal's choice of contracts and enforcement capacity, limiting how much we can learn from our model. We overcome this indeterminacy by dropping common knowledge of X and selecting a unique equilibrium. We do so by introducing noisy signals about X and taking the noise to zero following the global games approach of Frankel, Morris and Pauzner (2003).

Formally, we assume that each agent receives a signal $x = X + \nu\eta$, where $\nu > 0$ is a scaling factor and η is an i.i.d. random variable characterized by a continuous distribution H with full support on $[-1/2, 1/2]$. The signal is the only source of information about X. In particular, we assume that agents' prior about X is uniformly distributed on the interval $[0, 1]$.[14] Loosely speaking, the noise represents an agent's "uncertainty" or lack of perfect confidence in their inference about fundamentals, such as X or the loan contract.

We first show that there is a unique limit equilibrium, given by threshold strategies.

Proposition 1. *The enforcement game has a unique equilibrium as $\nu \to 0$.*[15] *Equilibrium strategies are characterized by a signal cutoff $k(w)$ such that*

- *If $x \geq k(w)$, agents choose to repay ($a(x) = 1$)*
- *If $x < k(w)$, agents choose to default ($a(x) = 0$).*

The reason why uniqueness obtains is that the introduction of small uncertainty about X induces large strategic uncertainty about the equilibrium actions of others (i.e., about the default

[14] Our results do not hinge upon the uniform prior assumption. As Frankel, Morris and Pauzner (2003) show, equilibrium selection arguments work in the limit as signal error goes to zero since any well-behaved prior will be approximately uniform over the small range of X that are possible given an agent's signal.

[15] Equilibrium strategies are unique up to sets of measure zero.

rate), hindering agents' ability to sustain multiple equilibria by seamlessly coordinating their beliefs about ψ.[16] The proof is fairly standard and roughly follows the logic in Frankel, Morris and Pauzner (2003).[17]

Although the previous result establishes the cutoff nature of equilibrium strategies, to characterize equilibrium, we need to pin down $k(w)$ in the limit as $\nu \to 0$. To do so, we need to solve the set of indifference conditions

$$E(P|x = k(w)) = \theta_{\bar{w}}(w) \text{ for all } w \text{ with } \theta_{\bar{w}}(w) \in (0, 1), \tag{11}$$

which state that an agent is indifferent between defaulting or not when she receives signal $x = k(w)$. This, however, becomes a challenging task when cutoffs are heterogeneous. To see why, note that to solve (11) we need to know the distribution of $P = \min\{X/\psi, 1\}$ conditional on $x = k(w)$. However, while the information about X is arbitrarily precise since $\nu \downarrow 0$, determining the conditional distribution of ψ involves pinning down agent strategic beliefs about the behavior of others, which are a complicated object when cutoffs vary with w.

To see that it is indeed heterogeneity that makes this approach intractable, let us look at what would happen if agents used a single threshold $k(w) = k$. In this simple case, the default rate of strategic agents (those with $\theta_{\bar{w}}$ between 0 and 1) is given by the fraction of those agents receiving signals below k. This implies that, since signal noise is i.i.d., an agent whose signal

[16] The presence of large strategic uncertainty even at infinitesimal noise levels rests on agents' higher order beliefs. Note that when an agent observes signal x, she considers it possible that X is $\nu/2$ away from her signal. As a result, she also admits that other agents may observe a signal as far as ν away from her own signal, and thus that they admit a possibility that X is as far as $\frac{3}{2}\nu$ away from her signal, thus placing positive probability that other agents think other agents observe signals as far as 2ν away from her signal. When this reasoning is repeated ad infinitum, it is clear that infinite order beliefs about X will fan out for any arbitrarily small ν. This divergence of higher order beliefs translates into divergent beliefs about ψ.

[17] The formal proof uses the fact that games with strategic complementarities feature a smallest and largest Nash equilibrium, both in cutoff strategies (Milgrom and Roberts, 1990), and shows that both must coincide. Specifically, it shows that, given the noise structure, expected monitoring probabilities conditional on receiving the cutoff signal are increasing in signal cutoffs, implying that there can be only one profile of cutoffs at which agents are indifferent between defaulting or not.

$x = X + \mu\eta$ is equal to the cutoff k believes that the strategic default rate is given by $H(\eta)$ regardless of ν (i.e., by the mass of agents with signal noise lower than hers). But since the agent does not observe η, she views $H(\eta)$ as a random variable, which is distributed uniformly in $[0, 1]$—the cdf of a random variable is uniformly distributed. Intuitively, when noise is i.i.d. she has no information about the ranking of her signal among all realized signals (given by $H(\eta)$), which is what determines the default rate when $x = k$. This feature of beliefs is known as the *Laplacian* property (Morris and Shin, 2003).

Unfortunately, the Laplacian property no longer holds when $k(w)$ is heterogeneous and solving for (11) as $\nu \downarrow 0$ looks hopeless. The reason is that agents with different cutoffs exhibit different default rates. This is not necessarily a problem when solving for $E(P|x = k(w))$ when $k(w)$ is not within ν of the other thresholds: Since other agents' signals fall into $[x - \nu, x + \nu]$, the agent knows exactly the default rate when $x = k(w)$ and can infer P as signal noise vanishes. However, this is troublesome for agents with thresholds within ν of each other, i.e., those whose thresholds converge to the same limit as $\nu \downarrow 0$ and hence hold nondegenerate beliefs about ψ. Moreover, we do not know when such a clustering may occur without solving for (11), and to solve these conditions we must know the distribution of ψ conditional on $x = k(w)$.

Despite this, we are able to determine equilibrium thresholds by adapting a result established in the context of a simpler model by Sakovics and Steiner (2012), which shows that although the Laplacian property does not apply to any individual type, it still applies on *average*. We use this remarkable result to circumvent the need to pin down individual beliefs and analytically characterize equilibrium thresholds by *averaging* the indifference conditions of types in any potential cluster and using them to identify which cluster arises in equilibrium and its corresponding signal threshold. That is, we use the average condition

$$\int_{w \in W'} E(P|x = k(w))dF(w) = \int_{w \in W'} \theta_{\bar{w}}(w)dF(w), \tag{12}$$

where W' is a subset of types that may cluster on the same limit threshold $k(w)$ as $\nu \downarrow 0$, and replace individual beliefs about ψ by the average belief in the cluster. Since $X = k(w)$ in the limit when $x = k(w)$, we can characterize the average expectation in (12) and pin down $k(w)$.

Specifically, Sakovics and Steiner (2012) establish that if we average agents' beliefs about ψ when they receive their threshold signal $x = k(w)$, weighted by their presence in the population $(f(w))$, we obtain the uniform distribution. In other words, if we randomly select a sample of agents and ask them for their beliefs about the distribution of ψ when they receive their threshold signal, the average answer is the uniform distribution. They call this property the *belief constraint*. Roughly speaking, it is driven by the fact that the beliefs in higher default rates conditional on $x = k(w)$ held by agents with high $k(w)$ are offset by the low-$k(w)$ agents' beliefs in lower default rates. Although Sakovics and Steiner (2012) derive the belief constraint by averaging the beliefs of the whole population to characterize the single limit threshold that emerges in games with symmetric equilibria, it generalizes to any subset of types in the game. As a consequence, it can be used in environments that yield asymmetric equilibria, as it is not necessary to know a priori which agent types are going to cluster in the limit.

Lemma 3 (belief constraint)**.** *Let $\psi(W', X)$ be the proportion of agents in a measurable set $W' \subseteq W$ choosing $a = 0$ when capacity is X, i.e.,*

$$\psi(W', X) := \frac{1}{\int_{W'} dF(w)} \int_{W'} H\left(\frac{k(w) - X}{\nu}\right) dF(w).$$

Then, for any $z \in [0, 1]$,

$$\frac{1}{\int_{W'} dF(w)} \int_{W'} \mathbb{P}_w \left(\psi(W', X) \leq z \middle| x = k(w)\right) dF(w) = z, \tag{13}$$

where $\mathbb{P}_w \left(\cdot \middle| x = k(w)\right)$ is the probability assessment of $\psi(W', X)$ by an agent receiving $x = k(w)$.

The belief constraint is instrumental in characterizing equilibrium thresholds in our model as ν goes to zero. It allows us to express the average indifference condition (12) of types in cluster W' in a closed form and to identify any clusters that may arise in equilibrium. The next proposition presents the resulting equilibrium threshold, which is closely related to the shape of $\theta_{\bar{w}}(w)F(w)$. This function represents the level of X that makes agents of type w indifferent between defaulting or not when X is common knowledge and all agents with returns lower than w default ($\psi = F(w)$).[18] In this regard, Assumption 1 leads to the existence of a unique cluster by ensuring that $\theta_{\bar{w}}(w)F(w)$ is single peaked, although our characterization technique applies to settings with multiple thresholds (see Theorem 2 in the Appendix).

Lemma 4. *If Assumption 1 holds then $\theta_{\bar{w}}(w)F(w)$ is single peaked and is increasing at 0.*

Let $w_{\max}$=argmax$_w\, \theta_{\bar{w}}(w)F(w)$.

Proposition 2. *In the limit, as $\nu \to 0$, there exists $w^* \geq \bar{w}$ such that*

$$k(w) = \begin{cases} \theta_{\bar{w}}(w)F(w) & \text{for all } w > w^* \\ \theta_{\bar{w}}(w^*)F(w^*) & \text{for all } w \in [\bar{w}, w^*] \end{cases}$$

where $w^* = \bar{w}$ if $\bar{w} > w_{\max}$ and, when $\bar{w} < w_{\max}$, w^* is the unique solution in $(w_{\max}, \infty)$ to

$$\theta_{\bar{w}}(w^*)F(w^*)\left(1 - \log \theta_{\bar{w}}(w^*)\right) - F(\bar{w}) = \int_{\bar{w}}^{w^*} \theta_{\bar{w}}(w)dF(w). \tag{14}$$

Furthermore, $\theta_{\bar{w}}(w^*)F(w^*)$ is increasing in $\bar{w}$.

Figure 1 illustrates the equilibrium strategy implied by Proposition 2. It plots $k(w)$ in relation to $\theta_{\bar{w}}(w)F(w)$. As is clear from the figure, $k(w)$ is equal to $\theta_{\bar{w}}(w)F(w)$ for $w > w^*$. However,

[18] Under common knowledge of X, ψ is common knowledge in equilibrium. Hence, when $\psi = F(w)$, indifference implies $P^e = X/F(w) = \theta_{\bar{w}}(w)$.

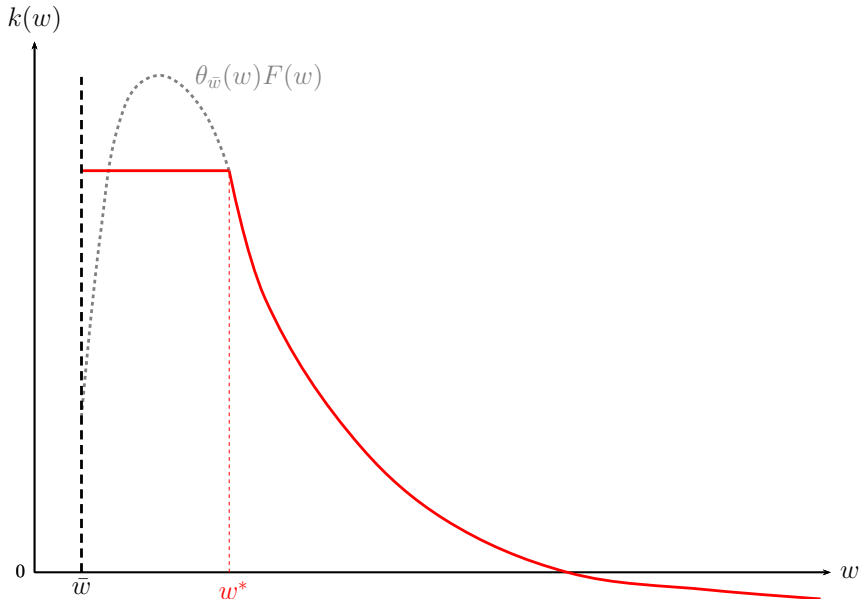

Figure 1: Equilibrium Strategy in the Global Game

this is not the case on the interval $w \in [\bar{w}, w^*]$, where agents from the interval follow the same threshold strategy $k(w) = k(w^*)$. Consequently, equilibrium features a bang-bang property. That is, at sufficiently high X, the equilibrium involves the lowest possible default rate, $F(\bar{w})$. However, if X falls below $\theta_{\bar{w}}(w^*)F(w^*)$ the default rate discontinuously jumps because agents with types between w and w^* simultaneously decide to default strategically.

Clustering arises in our model despite heterogeneity in default propensities due to a contagion effect. As signal noise vanishes, if agents with higher propensity $\theta_{\bar{w}}$ use a higher signal threshold than agents with slightly lower $\theta_{\bar{w}}$, when the latter receive their threshold signals they are certain that the higher propensity types are defaulting. If these high-$\theta_{\bar{w}}$ types have a strong presence in the population, such an increase in the default rate is enough to push the expected monitoring probability of the lower propensity agents below their respective $\theta_{\bar{w}}$, inducing them to default. That is, they would rather use the same signal threshold as the high-propensity types. This effect snowballs down the distribution of types until subsequent types have a weaker presence in the population so that higher default rates no longer compensate for their lower propensity

to default. At that point, signal thresholds become strictly decreasing.

Formally, the condition for the presence of a (single) cluster is $\theta_{\bar{w}}(w)F(w)$ being single peaked, guaranteed by Assumption 1. To see why notice that, if cutoffs were strictly decreasing at all w, an agent receiving her threshold signal is certain that the default rate is $F(w)$, leading to indifference condition $k(w) = \theta_{\bar{w}}(w)F(w)$, which is single peaked rather than decreasing. This implies that F initially goes up fast enough with w to more than compensate for the decrease in $\theta_{\bar{w}}(w)$. Accordingly, the only way to satisfy indifference conditions is to have a cluster containing agent types on both sides of the peak. Note that multiple peaks due to a multimodal distribution of returns or to nonmonotonic default propensities can lead to several clusters. Theorem 2 in the Appendix shows how to characterize them.

A remaining question is that if all agents in the cluster use the same threshold, how is it possible to satisfy their heterogeneous indifference conditions simultaneously? The answer is that away from the limit ($\nu > 0$), thresholds of types in the cluster are heterogeneous but are within ν of each other so that different types still have different expected monitoring probabilities conditional on receiving their threshold signals.

Finally, the last part of Proposition 2 establishes that the cluster's threshold goes up with repayment cutoff $\bar{w}$. This is due to two effects: a *direct effect* by increasing the default propensities of all agents, and an *externality effect* by increasing the pool of strategic agents in the population (recall that the range of strategic agents is given by $(\bar{w}, \bar{w}/(1-\gamma\mu))$ by Lemma 2).

3.2 Credit provision

We now turn to the contracting stage after X has been fixed. The principal's goal at this stage is to select $(b, \bar{w})$ to maximize entrepreneurs' net payoff subject to zero profit condition (10) and to the repayment behavior determined by capacity X and equilibrium thresholds $k(w)$. To do so, we first establish that the principal's problem has an interior solution.

Lemma 5. *The solution to the principal's problem* (9)-(10) *involves finite* b *for all* $X \in [0,1]$.

The next proposition formally establishes that credit levels are increasing in X whenever enforcement externalities "bite", in other words, when it is optimal to set b so that X prevents the cluster from defaulting. This is the case if X is not too low so that averting strategic default would require tiny loan levels, resulting in a cluster so small that it is better to provide slightly higher credit and let the cluster default.[19]

Proposition 3. *If* $X \geq \theta_{\bar{w}}(w^*)F(w^*)$ *at the optimal contract then* b *increases with* X.

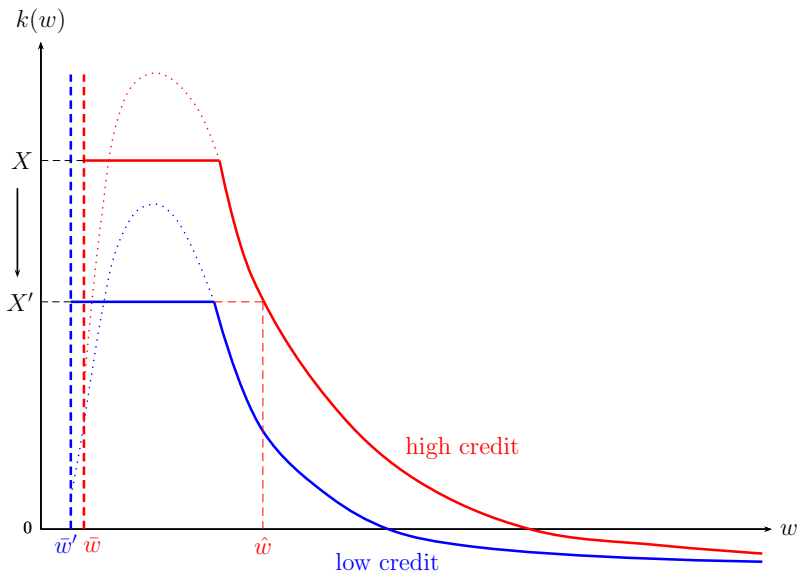

Figure 2: Equilibrium Strategy in the Global Game.

Figure 2 depicts this result. At a high X, b is high because there is enough capacity to sustain a high repayment cutoff $\bar{w}$ while enforcing a low default rate $F(\bar{w})$. At X', however, the previous default rate $F(\bar{w})$ becomes unsustainable for the same value of b. In fact, had the contract remained unchanged, a segment of agents between $\bar{w}$ and $\hat{w}$ would default. Since

[19] See footnote 13.

default is socially wasteful, the principal prefers to lower b to avert a default wave. As a result, both investment and output fall in the economy.

A key implication is that as long as letting agents in the cluster default is inefficient, a shock that eats up enforcement capacity before contracts are issued necessarily leads to credit rationing in the new loans market. That is, shocks to enforcement capacity can generate a credit crunch. Similarly, a distortion in the objective function of the principal reflecting political economy frictions that raise α, results in less capacity buildup and, as a consequence, in lower credit provision. We next consider two numerical examples inspired by the literature that illustrate these effects.

4 Applications

The first application concerns the propagation of macroeconomic shocks. The second application deals with the impact of political economy distortions on a developing economy that is in the process of accumulating enforcement capacity to deepen its financial markets. In both applications, we consider the full equilibrium of our model. In other words, the solution involves the endogenous choice of enforcement capacity X and credit contracts.

4.1 Propagation of macroeconomic shocks

In this example, the question we are interested in exploring is how a single shock of size $s > 0$ affects credit provision. We focus on the baseline timing (i.e., the shock hits and is observed by the principal before the contracting stage) but also discuss the possible outcomes under the alternative timing in which the shock hits after loans have been issued.

To set up this exercise, we assume that R is high enough so that the nonnegativity constraint on the principal's own consumption $g \geq 0$ does not bind. Hence, the principal chooses X solely

based on her preferences and on the probability of the shock $s > 0$. We assume that her preferences are undistorted in the sense that the principal values ex ante resources on par with entrepreneur utility (i.e. $\alpha = 1$). The value of γ is arbitrary and set equal to .25. A higher value of γ considerably strengthens our results, while a lower value weakens them, with $\gamma = 0$ corresponding to a frictionless benchmark under which the shock has no bite since there is no need for the principal to accumulate capacity.

Table 1: Parameters and Aggregate Statistics

Parameter	Value	Source
y	1	
Ew	1.02	BGG[a]
F	Lognormal ($\sigma = 3/8$)	BGG, Christiano et al. (2014)[b]
μ	0.88	BGG
γ	0.25	
$c(X)$	$0.088X$	

Statistic	Value	Target
Debt-to-equity ($\frac{b}{y}$)	0.8	0.5 - 1[c]
Default rate (ψ)	2.3%	2.3%[d]
Return on equity (ROE)	3.4%	
Capacity costs/ROE	0.06	
Cluster size ($Pr(w \in [\bar{w}, w^*])$)	1.7%	
% strategic agents ($\theta_{\bar{w}}(w) \in (0,1)$)	6.9%	

[a] BGG refers to Bernanke et al. (1999).
[b] The std. deviation of $log(w)$ is between 1/4 (Christiano et al., 2014) and 1/2 (BGG). It is chosen, along with $c(X)$, to match the leverage and default rate targets.
[c] Christiano et al. (2014) set the debt-to-equity ratio of 0.52, while BGG have it equal to 1.
[d] 2000-2007 Average delinquency rate on business loans. Source: Board of the Federal Reserve.

In terms of other parameters, we set them so that our model is consistent with data targets used in the quantitative literature studying the effect of financial frictions, in particular with the ones used in financial accelerator models such as Bernanke et al. (1999) and Christiano et al. (2014). Table 1 lists our parameter choices and calibration targets.

We assume the technology to accumulate capacity is linear and set unit costs so that the model delivers a debt-to-equity ratio of 0.8 and a default rate of 2.3% in the absence of the shock, which is the average delinquency rate on commercial loans in the U.S. from 2000 to 2007. In addition, our model implies a sensible level of equity premium of about 3.4% (after tax). Our choice of unit costs translates into aggregate capacity costs equivalent to 6% of the aggregate return on equity in the economy. Finally, we set the size of the shock to $s = 0.018$ to capture the magnitude of the recent financial crisis. Specifically, we use the fact that the default rate on commercial loans went up from an average of about 2.3% before the crisis to 4.1% in 2009.[20] Although we look at credit provision for different shock probabilities, our calibration assumes that the shock hits with low probability (1%).

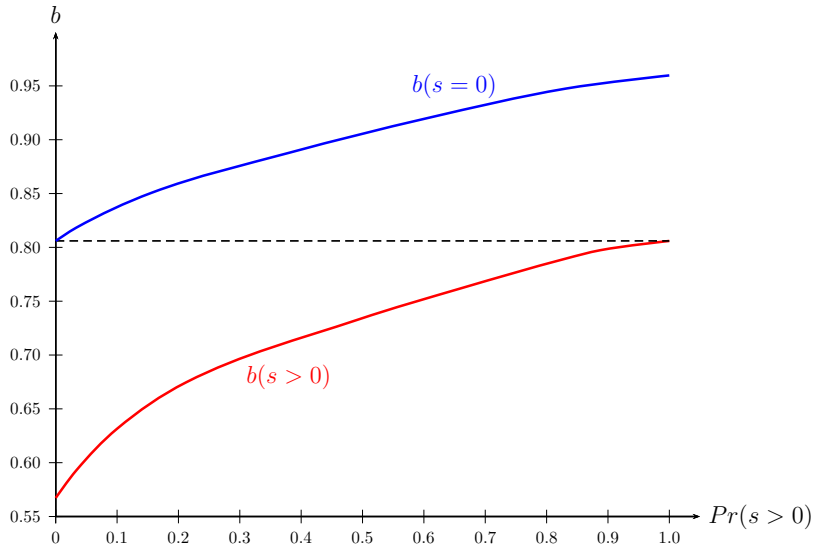

Figure 3: Enforcement Capacity and Shock Propagation

Figure 3 illustrates the impact of the shock on credit. It plots loan levels for each realization of the shock as a function of the shock probability. When the probability of the shock is very low, the shock implies a contraction of credit by 30% (relative to the frictionless benchmark

[20]Source: Board of Governors of the Federal Reserve System.

of $\gamma = 0$). Interestingly, although the gap $b(s = 0) - b(s > 0)$ shrinks as the shock becomes more likely, it is still sizable at all shock probabilities. The reason behind this surprising result is that higher shock probabilities induce the principal to accumulate precautionary capacity X to dampen the effect of the shock on credit provision, but this extra capacity leads to an "oversupply" of credit in normal times that sustains the gap.

These results suggest that enforcement externalities can play a crucial role in propagating large macroeconomic shocks that increase the rate of nonperforming loans in the economy, even when there is ample precautionary capacity to deal with them. Importantly, they are not driven by an unrealistically high presence of agents who would default for strategic reasons: In our calibration just about 7% of agents would consider strategic default, and our numerical results are driven by the fact that about 25% of them, or 1.7% of agents in the population, cluster on the same equilibrium strategy.

Alternative timing. Under the alternative timing of events, we assume that the principal signs a noncontingent contract $(b, \bar{w})$ before the aggregate shock realization occurs. The shock can be either anticipated or unanticipated. If the aggregate shock is unanticipated, our model implies that an entire pool of agents between $\bar{w}$ and $\hat{w}$ in Figure 1 defaults, and the principal violates her resource feasibility constraint (10). In this context, the principal, or lenders in the case of a competitive loan market, either default on their obligations or must be bailed out by external creditors, as in the sovereign crisis model of Arellano and Kocherlakota (2014). In our calibrated example, an unanticipated shock would lead to a jump in the default rate from 2.3% to more than 4%, depending on the size of the shock.

The effect of an anticipated shock under the alternative timing depends on whether the equilibrium in which agent types between $\bar{w}$ and w^* default is feasible after the shock hits. If it is feasible – and it may well not be when the cluster is large enough – the principal has the

option of letting all agents in the cluster default. The benefit of doing so is that the principal does not need to cut on b when $s = 0$, while the cost is the deadweight loss associated with a wave of coordinated defaults when $s > 0$. This implies that ex post inefficient crises might be ex ante efficient in our environment depending on the size and likelihood of the shock. If a wave of defaults under the shock is not ex ante optimal, the economy will suffer from inefficiently low credit provision in both normal and distressed times.

4.2 Propagation of political economy frictions

Our second application concerns an economywide planning problem that involves accumulation of state capacity X so as to relax credit constraints and grow the economy. In contrast to the first application, there are no shocks, and R is such that the constraint $g \geq 0$ may be binding at some values of γ and α. That is, the economy is short of resources and the principal would like to relax credit constraints.

We use this example to highlight that in the presence of strong enforcement externalities, credit provision may respond in a highly nonlinear way to political economy distortions captured by preference parameter α. To do so, we set parameter values similar to the previous setup except for R and solve the model for different values of α and γ. Specifically, we set $R = 0.0045$, which implies that $g \geq 0$ is just binding at $\gamma = 0.25$ and $\alpha = 0$ – recall that $\alpha = 0$ represents a principal who only cares about entrepreneurs' utility and hence uses resources to build as much capacity as possible at the expense of alternative uses of resources (g).

The results are illustrated in Figure 4. The horizontal axis shows distortion level α, while the vertical axis reports the supply of credit. Clearly, for low values of α, the principal sets $g = 0$ and accumulates as much capacity as possible. Not surprisingly, as α increases, the planner diverts more resources for consumption, building less capacity and thus causing a reduction in

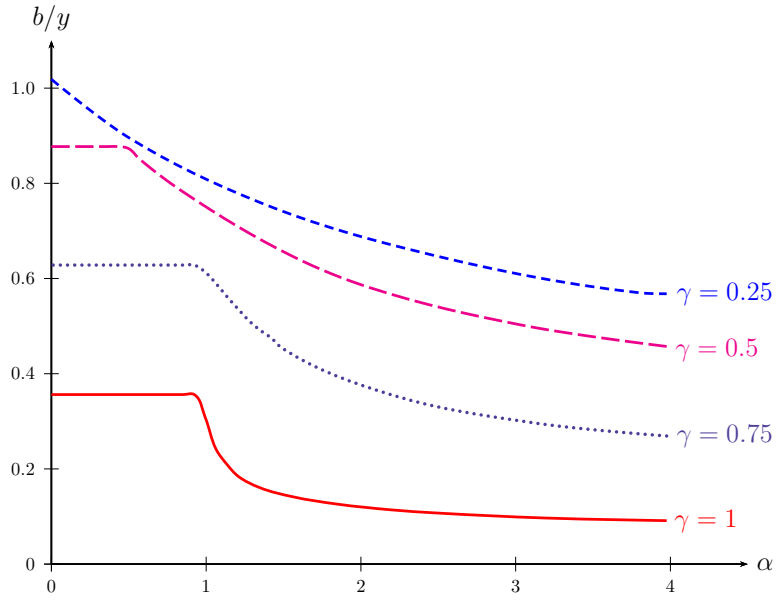

Figure 4: Credit in an Economy with Political Distortions

the credit supply. What is interesting in the result is how the relationship between α and b changes as enforcement externalities become stronger (i.e., as γ goes up). Whereas credit drops in a parsimonious fashion as α goes up at $\gamma = 0.25$, two effects kick in at high γ. First, the resource constraint binds at low α making credit unresponsive to small changes in α. This is because the pool of strategic agents is larger at higher γ (Lemma 2), and thus, higher levels of capacity are necessary to sustain a given credit supply. Second, when α is high enough to make the principal willing to start diverting resources away from capacity buildup ($g \geq 0$ is no longer binding), a small increase in α can cause severe reductions in the credit supply. This can lead to *financial development traps* characterized by the unraveling of credit markets in response to small changes in the opportunity costs of building state capacity. The reason for the steep reaction of credit to distortions is that the effect of an increase of the repayment cutoff $\bar{w}$ on the size of the strategic agent pool is larger at higher γ, thus requiring a bigger change in capacity to sustain such an increase in $\bar{w}$. Accordingly, b is less responsive to changes in capacity, causing the

marginal benefit of building capacity to go down as enforcement externalities become stronger.

The main implication of our results for the theory of state capacity proposed by Besley and Persson (2009, 2010) is that poor countries with inadequate enforcement institutions (i.e., those where externalities may be stronger) may be particularly exposed to development traps. The reason is that they have weak incentives to develop enforcement capacity, since initial investments in capacity are relatively ineffective in spurring credit and growth. From a policy perspective, our model suggests that development aid targeting a country's enforcement infrastructure may be critical in propping up its credit markets.

Beyond the context of this particular example, our theory provides an alternative microfoundation of state capacity that highlights the endogeneity of contract enforceability. According to this view, adequate enforcement institutions are needed to make sure that default spillovers do not lead to the unraveling of credit markets.

5 Conclusions

We have analyzed the effect of capacity-constrained enforcement in a standard model of debt financing and devised new methods to study the effect of this friction in the presence of agent heterogeneity. Our results suggest that heterogeneous agents may partially pool on the same equilibrium strategy, implying far less heterogeneity in terms of equilibrium behavior and thus leading to macroeconomic fragility. We have shown that enforcement externalities can significantly impact the functioning of credit markets by considering two distinct applications. In the first application, we have shown that our model can explain two key characteristics of financial crises: disproportionate jumps in default rates and credit crunches that are resolved only after widespread liquidation of bad assets takes place. Interestingly, we have demonstrated that contraction of credit may occur even in economies in which there is significant amount of pre-

cautionary capacity accumulation to preempt such shocks. Our second application contributes to the theory of the origins of state capacity (Besley and Persson, 2009, 2010) by highlighting how enforcement externalities influence a country's incentives to build market-supporting institutions and exacerbate the effect of political economy distortions on economic development.

References

Arellano, Cristina and Narayana Kocherlakota, "Internal Debt Crises and Sovereign Defaults," *Journal of Monetary Economics*, 2014, *68*, S68–S80.

Bassetto, Marco and Christopher Phelan, "Tax Riots," *Review of Economic Studies*, 2008, *75* (3), 649–669.

Bernanke, Ben S., Mark Gertler, and Simon Gilchrist, "The Financial Accelerator in a Quantitative Business Cycle Framework," *Handbook of Macroeconomics*, 1999, *1*, 1341–1393.

Besley, Timothy and Torsten Persson, "The Origins of State Capacity: Property Rights, Taxation, and Politics," *American Economic Review*, September 2009, *99* (4), 1218–1244.

_ **and** _ , "State Capacity, Conflict, and Development," *Econometrica*, 2010, *78* (1), 1–34.

_ **and** _ , *Pillars of Prosperity: The Political Economics of Development Clusters*, Princeton University Press, 2011.

Bond, Philip and Ashok S. Rai, "Borrower Runs," *Journal of Development Economics*, 2009, *88* (2), 185–191.

_ **and Kathleen Hagerty**, "Preventing Crime Waves," *American Economic Journal: Microeconomics*, 2010, *2* (3), 138–159.

Calem, Paul, Julapa Jagtiani, and William W Lang, "Foreclosure Delay and Consumer Credit Performance," *Federal Reserve Bank of Philadelphia Working Paper*, 2015.

Carlsson, Hans and Eric van Damme, "Global Games and Equilibrium Selection," *Econometrica*, September 1993, *61* (5), 989–1018.

Carlstrom, Charles T., Timothy S Fuerst, and Matthias Paustian, "Optimal Contracts, Aggregate Risk, and the Financial Accelerator," *American Economic Journal: Macroeconomics*, 2016, *8* (1), 119–147.

Carrasco, Vinicius and Pablo Salgado, "Coordinated Strategic Defaults and Financial Fragility in a Costly State Verification model," *Journal of Financial Intermediation*, 2014, *23* (1), 129–139.

Chan, Sewin, Andrew Haughwout, Andrew Hayashi, and Wilbert Van der Klaauw, "Determinants of Mortgage Default and Consumer Credit Use: The Effects of Foreclosure Laws and Foreclosure Delays," *Federal Reserve Bank of New York Staff Report*, 2015.

Christiano, Lawrence J., Roberto Motto, and Massimo Rostagno, "Risk Shocks," *American Economic Review*, 2014, *104* (1), 27–65.

Cordell, Larry, Liang Geng, Laurie Goodman, and Lidan Yang, "The Cost of Delay," *Federal Reserve Bank of Philadelphia Working Paper*, 2013.

Diamond, Douglas W and Philip H Dybvig, "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, 1983, *91* (3), 401–419.

Djankov, Simeon, Oliver Hart, Caralee McLiesh, and Andrei Shleifer, "Debt Enforcement Around the World," *Journal of Political Economy*, 2008, *116* (6), 1105–1149.

Enoch, Charles, Barbara Baldwin, Frécaut Olivier, and Arto Kovanen, "Indonesia: Anatomy of a Banking Crisis. Two Years of Living Dangerously, 1997-99.," *IMF Working Paper*, 1998, (0152).

Frankel, David M., Stephen Morris, and Ady Pauzner, "Equilibrium Selection in Global Games with Strategic Complementarities," *Journal of Economic Theory*, 2003, *108* (1), 1–44.

Gale, Douglas and Martin Hellwig, "Incentive-Compatible Debt Contracts: The One-Period Problem," *Review of Economic Studies*, 1985, *52* (4), 647–663.

Herkenhoff, Kyle F. and Lee Ohanian, "Foreclosure Delay and US Unemployment," *Federal Reserve Bank of St. Louis Working Paper Series*, 2012.

Iverson, Benjamin, "Get in Line: Chapter 11 Restructuring in Crowded Bankruptcy Courts," *Working Paper*, 2015.

Jappelli, Tullio, Marco Pagano, and Magda Bianco, "Courts and Banks: Effects of Judicial Enforcement on Credit Markets," *Journal of Money, Credit, and Banking*, 2005, *37* (2), 223–244.

Krasa, Stefan and Anne P. Villamil, "Optimal Contracts when Enforcement Is a Decision Variable," *Econometrica*, 2000, *68* (1), 119–134.

Krueger, Anne and Aaron Tornell, "The Role of Bank Restructuring in Recovering from Crises: Mexico 1995-98," *NBER Working Paper*, 1999.

Mayer, Christopher, Edward Morrison, Tomasz Piskorski, and Arpit Gupta, "Mortgage Modification and Strategic Behavior: Evidence from a Legal Settlement with Countrywide," *American Economic Review*, 2014, *104* (9), 2830–2857.

Milgrom, Paul and John Roberts, "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities," *Econometrica*, 1990, *58* (6), 1255–1277.

Morris, Stephen and Hyun Song Shin, "Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks," *American Economic Review*, 1998, *88* (3), pp. 587–597.

_ **and** _ , "Global Games: Theory and Applications," in "Proceedings of the Eighth World Congress of the Econometric Society," Cambridge University Press, 2003.

Ponticelli, Jacopo, "Court Enforcement and Firm Productivity: Evidence from a Bankruptcy Reform in Brazil," *Quarterly Journal of Economics*, forthcoming.

Safavian, Mehnaz and Siddharth Sharma, "When Do Creditor Rights Work?," *Journal of Comparative Economics*, 2007, *35* (3), 484–508.

Sakovics, Jozsef and Jakub Steiner, "Who Matters in Coordination Problems?," *American Economic Review*, 2012, *102* (7), 3439–3461.

Shleifer, Andrei and Robert Vishny, "Fire Sales in Finance and Macroeconomics," *Journal of Economic Perspectives*, 2011, *25* (1), 29–48.

Stone, Mark R., "Large-Scale Post-Crisis Corporate Sector Restructuring," *IMF Policy Discussion Paper*, 2000.

Townsend, Robert, "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory*, 1979, *21* (2), 265–293.

Vives, Xavier, "Nash Equilibrium with Strategic Complementarities," *Journal of Mathematical Economics*, 1990, *19* (3), 305–321.

Woo, David, "Two Approaches to Resolving Nonperforming Assets During Financial Crises," *IMF Working Paper*, 2000.

Appendices

Appendix I: Equilibrium multiplicity under common knowledge

In this Appendix, we show that under common knowledge of X, our model exhibits multiple equilibria in a range of X. Under common knowledge, the equilibrium ψ and P are also common knowledge. Hence, by Lemma 1, an agent of type w defaults if $P < \theta_{\bar{w}}(w)$ and repays otherwise. Since default propensities are strictly decreasing in w, the equilibrium default rate is given by $\psi = F(\hat{w})$, where $\hat{w}$ is the agent type that is indifferent between defaulting and repaying. Accordingly, $\hat{w}$ solves $P = \theta_{\bar{w}}(\hat{w})$, and the equilibrium monitoring probability is given by

$$P^e = P = \min\left\{\frac{X}{F(\hat{w})}, 1\right\}. \tag{15}$$

Unless the value of X is very high or very low, three equilibria are possible. In the first equilibrium, only the *truly* insolvent agents default, that is, agents with $w < \bar{w}$. In the additional two equilibria, some solvent agents default strategically.

The efficient equilibrium requires $X \geq F(\bar{w})$. Intuitively, this must be the case because agent types with w above but close to $\bar{w}$ have very little incentive to repay and requires $P = 1$ to prevent them from defaulting. In the other two equilibria, some agents with $w > \bar{w}$ default. These equilibria are sustained by a self-fulfilling belief that the capacity constraint binds and that monitoring is imperfect ($P < 1$). Otherwise, no agent with $w > \bar{w}$ would have an incentive to default.

These equilibria are pinned down by the above indifference condition $P = \theta_{\bar{w}}(\hat{w})$ and a binding capacity constraint $P = X/F(\hat{w})$, which lead to equation:

$$X = \theta_{\bar{w}}(\hat{w})F(\hat{w}), \tag{16}$$

This equation admits up to two solutions. This is because, under Assumption 1, $\theta_{\bar{w}}(w)F(w)$ is single peaked (see Lemma 4 below). Figure 5 illustrates the existence of the two inefficient equilibria exhibiting default rates $F(w_2)$ and $F(w_3)$, respectively. Proposition 4 formalizes this result.

Definition 1. *Let* $\underline{X} = F(\bar{w})$ *and* $\overline{X} = \theta_{\bar{w}}(w_{\max})F(w_{\max})$.

Proposition 4. *Under common knowledge of* X*, if* $\bar{w} < w_{\max}$*, equilibrium is unique iff* $X < \underline{X}$ *or* $X > \overline{X}$*. Otherwise, there are three equilibria in the case of* $X \in (\underline{X}, \overline{X})$ *and two equilibria in the case of* $X = \underline{X}$ *and* $X = \bar{X}$*. If* $\bar{w} \geq w_{\max}$ *equilibrium is always unique.*

Appendix II: Proofs

To prove the results, we proceed as follows. First, we present equilibrium existence, selection, and characterization results for the model with general discrete distribution of returns. Then, we provide the proofs of the results in the paper by deriving the implications of these results

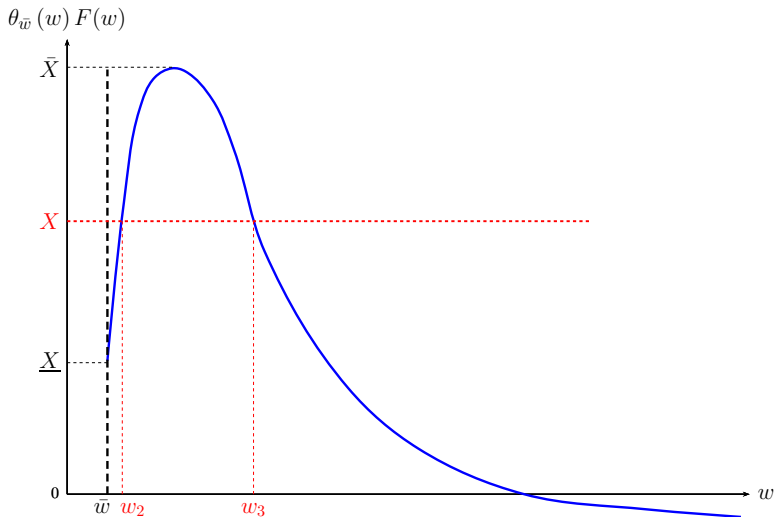

Figure 5: Multiplicity of Equilibria

when the distribution of returns is assumed to arbitrarily finely approximate some continuous distribution F that obeys Assumption 1.

A. General discrete distribution of returns

In this economy, $\mathcal{W} \subset [0, \infty)$ is a finite set of possible returns, each with a positive mass, which are distributed according to the commonly known discrete distribution F, with probability mass function f. Since the contract is fixed at the enforcement stage, we drop the subscript from $\theta_{\bar{w}}$ to lighten notation.

Given contract $(b, \bar{w})$, we make the following assumption about agent payoffs.

Assumption 3. *For all* $w \in \mathcal{W}$

(i) $U(0, w, 0) \neq U(1, w, 0)$*; and*

(ii) $U(0, w, 1) \neq U(1, w, 1)$.

Condition (i) implies that no agent is indifferent between paying back the loan and defaulting when $P = 0$, i.e., there is no agent with $\theta(w) = 0$. Similarly, (ii) means that that no agent is indifferent at $P = 1$, that is, there is no agent with $\theta(w) = 1$. This technical assumption simplifies the proof of uniqueness by implying the existence of dominance regions for ν sufficiently small.

Note that agents with $w < \bar{w}$ $(\theta(w) > 1)$ and those with $\theta(w) < 0$ behave in a nonstrategic fashion: The former always choose $a = 0$, and the latter $a = 1$, regardless of P. Hence, our focus is on pinning down the behavior of types in the set $\mathcal{W}^* := \{w \in \mathcal{W} : \theta(w) \in (0, 1)\}$, with its lowest and highest elements respectively denoted w_l and w_h.

We assume that the noise scale factor satisfies $0 < \nu < \bar{\nu} := \min\{\theta(w_h)F(\bar{w}), 1 - \theta(w_l)\}$.[21]

[21] This upper bound on ν is helpful to show uniqueness of equilibrium by ensuring that boundary issues associated with signals close to 0 or 1 only arise when capacity is such that all agents have a dominant strategy.

We first establish that there exists a unique equilibrium of the game with finite types, featuring cutoff strategies.

Theorem 1. *The game has an essentially unique equilibrium.*[22] *Equilibrium strategies are characterized by cutoffs $k(w)$ on signal x, such that all agents of type $w \in \mathcal{W}^*$ choose action $a = 1$ if $x \geq k(w)$ and $a = 0$ otherwise.*

Proof. The proof logic is as follows. First, we argue that the set of equilibrium strategy profiles has a largest and a smallest element, each involving monotone strategies (cutoff) strategies. Second, we show that there is at most one equilibrium in monotone strategies (up to differences in behavior at cutoff signals). But this implies that the equilibrium is essentially unique.

The existence of a smallest and largest equilibrium profile in monotone strategies follows from existing results on supermodular games by Milgrom and Roberts (1990) and Vives (1990). Consider the game in which we fix the profile $\mathbf{x}$ of signal realizations and agents choose actions $\{0, 1\}$ after observing their own signals. It is straightforward to check that the game satisfies the conditions of Theorem 5 in Milgrom and Roberts (1990), which states that the game has a smallest and largest equilibrium. That is, there exist two equilibrium strategy profiles, $\underline{\mathbf{a}}(\mathbf{x})$ and $\bar{\mathbf{a}}(\mathbf{x})$ such that any equilibrium profile $\mathbf{a}(\mathbf{x})$ satisfies $\underline{\mathbf{a}}(\mathbf{x}) \leq \mathbf{a}(\mathbf{x}) \leq \bar{\mathbf{a}}(\mathbf{x})$. Moreover, if we fix the action profile of all agents, the difference in expected payoff from choosing $a = 0$ versus $a = 1$ for any given agent is increasing in $\mathbf{x}$ since default rates are the same across signal profiles, while X is higher in expectation the higher the signal profile is, thus implying a higher expected monitoring probability. That is, expected payoffs exhibit increasing differences w.r.t. $\mathbf{x}$, and Theorem 6 in Milgrom and Roberts (1990) applies: $\underline{\mathbf{a}}(\mathbf{x})$ and $\bar{\mathbf{a}}(\mathbf{x})$ are nondecreasing functions of $\mathbf{x}$. But because an agent's strategy can only depend on her own signal, all agents must be following cutoff strategies.

To show that there is at most one equilibrium in monotone strategies, we make use of the following two lemmas. The first one shows that equilibrium cutoffs are bounded away from zero and one. The second lemma uses these bounds to establish the following translation result: When all cutoffs are shifted by the same amount Δ expected monitoring probabilities go up. Equipped with such results we will show how as we move from the smallest to the largest equilibria monitoring probabilities go up, implying that there must be a unique profile of cutoffs at which indifference conditions (17) are satisfied.

Let $\mathbf{k} + \mathbf{\Delta} = (k(w) + \Delta)_{w \in \mathcal{W}^*}$, while $\underline{\mathbf{k}}$ and $\bar{\mathbf{k}}$ represent the profile of cutoffs associated to the smallest and largest equilibrium, respectively. Abusing notation, let $\mathbb{E}[P|\mathbf{k}; x]$ represent the expected monitoring probability of an agent receiving signal x when agents use cutoff profile $\mathbf{k}$.

Lemma 6. *If $\mathbf{k}$ is a profile of equilibrium cutoffs then $k(w) \in [(\theta(w) - \nu/2)F(\bar{w}), \theta(w) + \nu/2]$ for all $w \in \mathcal{W}^*$.*

Proof. Note that k is an equilibrium if it solves the following set of indifference conditions:

$$\mathbb{E}_\theta[P|\mathbf{k}; k(w)] = \theta(w) \qquad \forall w \in \mathcal{W}^*. \tag{17}$$

[22] In the sense that equilibrium strategies may differ in zero probability events.

Note also that the value of X conditional on $x \in [\nu/2, 1-\nu/2]$ is at least $x - \nu/2$. Given this and the fact that monitoring probability is given by (15), we have that, if $k(w) \in [\nu/2, 1-\nu/2]$,

$$\mathbb{E}[P|\mathbf{k}; k(w)] \geq \mathbb{E}[X|\mathbf{k}; k(w)] \geq k(w) - \nu/2.$$

But this implies that $\mathbb{E}[P|\mathbf{k}; k(w)] > \theta(w)$ when $k(w) > \theta(w) + \nu/2$, a contradiction. A similar logic rules out $k(w) > 1 - \nu/2$, given that $\mathbb{E}[X|\mathbf{k}; x]$ is monotone in x and that $\theta(w_l) < 1 - \nu$. Likewise, when $k(w) \in [\nu/2, 1-\nu/2]$,

$$\mathbb{E}[P|\mathbf{k}; k(w)] \leq \mathbb{E}\left[\frac{X}{F(\bar{w})}\middle|\mathbf{k}; k(w)\right] \leq \frac{k(w) + \nu/2}{F(\bar{w})},$$

which, using a symmetric argument, yields the above lower bound on $k(w)$. □

Lemma 7. *If* $\mathbf{k}$ *is a profile of equilibrium cutoffs then* $\mathbb{E}[P|\mathbf{k}; k(w)] < \mathbb{E}[P|\mathbf{k} + \boldsymbol{\Delta}; k(w) + \Delta]$ *for all* $\Delta > 0$ *and all* $w \in \mathcal{W}^*$ *such that* $k(w) + \Delta \leq \bar{k}(w)$.

Proof. First note that the density of X conditional on an agent receiving signal $x \in [\nu/2, 1-\nu/2]$ is given by $h\left(\frac{x-X}{\nu}\right)$. Also notice that an agent of type w defaults if she receives a signal $x < k(w)$, and thus, the fraction of type-w agents defaulting when capacity is X is given by $H\left(\frac{k(w')-X}{\nu}\right)$. Since $\nu \leq \theta(w_h)F(\bar{w}) \leq theta(w)F(\bar{w})$ and, by Lemma 6, $k(w) \geq (\theta(w) - \nu/2)F(\bar{w})$, we have that $k(w) \geq \nu/2$. Likewise, $k(w) + \Delta \leq \bar{k}(w) \leq 1 - \nu/2$ by Lemma 6 and the fact that $\nu \leq 1 - \theta(w_l)$. Hence, we can obtain the following inequality by a well-defined change of variables:

$$\begin{aligned}
&\mathbb{E}[P(X)|\mathbf{k}; k(w)] = \\
&\int_{-1/2}^{1/2} \min\left\{\frac{X}{F(\bar{w}) + \sum_{w'} H\left(\frac{k(w')-X}{\nu}\right) f(w')}, 1\right\} h\left(\frac{k(w) - X}{\nu}\right) dX \\
&< \int_{-1/2}^{1/2} \min\left\{\frac{X + \Delta}{F(\bar{w}) + \sum_{w'} H\left(\frac{k(w')-X}{\nu}\right) f(w')}, 1\right\} h\left(\frac{k(w) - X}{\nu}\right) dX \\
&= \int_{-1/2}^{1/2} \min\left\{\frac{X'}{F(\bar{w}) + \sum_{w'} H\left(\frac{k(w')+\Delta-X'}{\nu}\right) f(w')}, 1\right\} h\left(\frac{k(w) + \Delta - X'}{\nu}\right) dX' \\
&= \mathbb{E}[P|\mathbf{k} + \boldsymbol{\Delta}; k(w) + \Delta].
\end{aligned}$$

The inequality is strict because $\mathbf{k}$ being an equilibrium profile means that $\mathbb{E}[P|\mathbf{k}; k(w)] = \theta(w) < 1$ for all $w \in \mathcal{W}^*$. Accordingly, monitoring probabilities, conditional on $x = k(w)$, are less than 1 for a positive measure of $X \in [x - \nu/2, x + \nu/2]$, and hence, expected monitoring probabilities go up strictly when capacity increases by Δ. □

Equipped with Lemma 7, we know argue that $\underline{\mathbf{k}} = \bar{\mathbf{k}}$. Assume, by way of contradiction, that

$\underline{k}(w) < \bar{k}(w)$ for some $w \in \mathcal{W}^*$. Denote $\hat{w} = \arg\max_{w \in \mathcal{W}^*}(\bar{k}(w) - \underline{k}(w))$ and $\hat{\Delta} = \bar{k}(\hat{w}) - \underline{k}(\hat{w})$. By Lemma 7, we have that

$$\theta(\hat{w}) = \mathbb{E}[P|\underline{\mathbf{k}}; \underline{k}(\hat{w})] < \mathbb{E}[P|\underline{\mathbf{k}} + \hat{\mathbf{\Delta}}; \bar{k}(\hat{w})] \leq \mathbb{E}[P|\bar{\mathbf{k}}; \bar{k}(\hat{w})] = \theta(\hat{w}),$$

where the last inequality comes from the fact that default rates at $\bar{\mathbf{k}}$ are lower than at $\underline{\mathbf{k}} + \hat{\mathbf{\Delta}} \geq \bar{\mathbf{k}}$, and thus, the expected monitoring probability conditional on $x = \bar{k}(\hat{w})$ is higher. □

Next, we proceed to characterize the limit equilibrium as $\nu \to 0$. First, note that for threshold profile $k(\cdot)$ to be an equilibrium profile, it must satisfy the set of indifference conditions (17).

To solve this system of equations, we need to pin down agent beliefs when they receive their threshold signals. To do so we make use of the full force of the *belief constraint* (Sakovics and Steiner, 2012): On average, conditional on $x = k(w)$, agents with types in a subset $W' \subseteq \mathcal{W}^*$ believe that the default rate of agents in W' is uniformly distributed in $[0, 1]$. The default rate in W' when capacity is X is given by

$$\psi(X, W') = \frac{1}{\sum_{W'} f(w)} \sum_{W'} H\left(\frac{k(w) - X}{\nu}\right) f(w). \tag{18}$$

Lemma 8 (belief constraint)**.** *For any subset $W' \subseteq \mathcal{W}^*$ and any $z \in [0, 1]$,*

$$\frac{1}{\sum_{W'} f(w)} \sum_{W'} \mathbb{P}\left(\psi(X, W') \leq z \middle| x = k(w)\right) f(w) = z. \tag{19}$$

Proof. The result follows directly from the proof of Lemma 1 in Sakovics and Steiner (2012). To see why, note first that Lemma 6 guarantees that threshold signals and thus the "virtual signals" defined in their proof fall in $[\nu/2, 1 - \nu/2]$, which is needed for their belief constraint to hold. Second, it is straightforward to check that all the arguments and results in their proof hold unmodified if we condition all the probability distributions used in the proof on the event $w \in \mathcal{W}'$ and focus on the aggregate action of agents with types in W' rather than the aggregate action in the population.[23] □

The previous result is instrumental in characterizing equilibrium thresholds as ν goes to zero. In particular, it allows us to derive closed-form solutions for the above indifference conditions from which we can obtain $\mathbf{k}$. In stating this result, we refer to a partition $\Phi = \{W_1, \cdots, W_I\}$ of $\mathcal{W}^*$ as being monotone if $\max W_i < \min W_{i+1}$, $i = 1, \cdots, I-1$, and denote the lowest and highest elements of W_i by $\underline{w}_i$ and $\bar{w}_i$, respectively. Also, let $F^-(w) = \sum_{w' < w} f(w')$.

Theorem 2. *In the limit, as $\nu \to 0$, the equilibrium cutoff strategies are given by a unique monotone partition $\Phi = \{W_1, \cdots, W_I\}$ and a unique vector $(k_1, \cdots, k_I)$ satisfying the following conditions:*

[23] When thresholds do not fall within ν of each other, the distribution $\tilde{F}$ of virtual errors $\tilde{\eta}$ need not be strictly increasing, and thus, its inverse may not be well defined. Defining $\tilde{F}^{-1}(u) = \inf\{\tilde{\eta} : \tilde{F}(\tilde{\eta}) \geq u\}$ takes care of this issue and ensures that the proof of Lemma 2 in Sakovics and Steiner (2012) applies to the general case.

(i) $k(w) = k(w') = k_i$ *for all* $w, w' \in W_i$.

(ii) $k_i > k_{i+1}$ *for all* $i = 1, \cdots, I-1$.

(iii) $\theta(\underline{w}_i)F^-(\underline{w}_i) \leq k_i \leq \theta(\bar{w}_i)F(\bar{w}_i)$ *for all* $i = 1, \cdots, I$.

(iv) $\int_{F^-(\underline{w}_i)}^{F(\bar{w}_i)} \min\left\{\frac{k_i}{z}, 1\right\} dz = \sum_{W_i} \theta(w)f(w)$ *for all* $i = 1, \cdots, I$.

Proof. From Theorem 1 we know that for each $\nu > 0$, there exists essentially a unique equilibrium, which is in monotone strategies. Let $k^\nu(w)$ represent the equilibrium threshold of type-w agents associated with $\nu > 0$, with $\mathbf{k}^\nu$ denoting the equilibrium cutoff profile. The first step of the proof is to show that $\mathbf{k}^\nu$ uniformly converge as $\nu \to 0$, and identify the set of indifference conditions that pin down the limit equilibrium. Let

$$A_w(z|\mathbf{k}^\nu, W') := \mathbb{P}\left(\psi(X, W') \leq z \middle| x = k^\nu(w)\right)$$

denote the strategic belief of an agent of type $w \in W'$ when she receives her threshold signal $x = k^\nu(w)$.

Lemma 9. *There exist a unique partition* $\{W_1, \cdots, W_I\}$ *and a set of thresholds* $k_1 > k_2 > \cdots > k_I$ *such that, as* $\nu \to 0$*, for all* $w \in W_i$*,* $i = 1, \cdots, I$*,* $k^\nu(w)$ *uniformly converges to* k_i*. Moreover, thresholds* $\mathbf{k} = (k_1, \cdots, k_I)$ *solve the system of limit indifference conditions*

$$\int_0^1 \min\left\{\frac{k_i}{F(\bar{w}) + \sum\limits_{\cup_{j<i}W_j} f(w') + z\sum\limits_{W_i} f(w')}, 1\right\} dA_w(z|\mathbf{k}, W_i) = \theta(w), \quad \forall w \in W_i,\ \forall i, \tag{20}$$

where $A_w(z|\mathbf{k}, W_i)$ *represents the strategic beliefs of type-*w *agents in the limit and satisfies the belief constraint* (19)*.*

See proof below.

Equipped with this set of indifference conditions, we next prove that the partition of types is monotone and that thresholds satisfy (iii) and (iv) in the theorem.

We show that the partition of types must be monotone by way of contradiction. Assume that there are two types $w > \hat{w}$ such that $w \in \mathcal{W}_i$ and $\hat{w} \in \mathcal{W}_m$ with $m > i$. First note that the LHS in (20) is bounded below by $\frac{k_i}{F(\bar{w}) + \sum\limits_{\cup_{j\leq i}W_j} f(w')}$ and bounded above by $\min\left\{\frac{k_i}{F(\bar{w}) + \sum\limits_{\cup_{j<i}W_j} f(w')}, 1\right\}$. Given this, since $\theta(\hat{w}) < 1$ the monitoring probability when all agents with types in W_m default is strictly less than 1, i.e., $\frac{k_m}{F(\bar{w}) + \sum\limits_{\cup_{j\leq m}W_j} f(w')} < 1$. Otherwise, (20) would be violated. In addition, $m > i$ implies that $k_m < k_i$ by the above lemma and that $\sum\limits_{\cup_{j<i}W_j} f(w') < \sum\limits_{\cup_{j\leq m}W_j} f(w')$. Combining

all this, we arrive at the following contradiction

$$\theta(w) \geq \min\left\{\frac{k_i}{F(\bar{w}) + \sum\limits_{\cup_{j<i} W_j} f(w')}, 1\right\} > \frac{k_m}{F(\bar{w}) + \sum\limits_{\cup_{j\leq m} W_j} f(w')} \geq \theta(\hat{w}).$$

The monotonicity of the type partition implies that $F(\bar{w}) + \sum\limits_{\cup_{j\leq i} W_j} f(w') = F(\overline{w}_i)$ and that $F(\bar{w}) + \sum\limits_{\cup_{j<i} W_j} f(w') = F^-(\underline{w}_i)$. Given this, it is straightforward to check that the above bounds on the LHS of (20) lead to condition (iii) in the theorem.

Finally, to obtain condition (iv) from (20), we make use of the belief constraint in the limit, which can be written as

$$\frac{1}{\sum\limits_{W_i} f(w)} \sum_{W_i} A_w(z|\mathbf{k}, W_i) f(w) = z. \tag{21}$$

Multiplying both sides of (20) by $\frac{f(w)}{\sum_{W_i} f(w)}$ and summing over $w \in W_i$ we get

$$\int_0^1 \min\left\{\frac{k_i}{F^-(\underline{w}_i) + z\sum\limits_{W_i} f(w')}, 1\right\} d\left(\frac{1}{\sum\limits_{W_i} f(w)} \sum_{W_i} A_w(z|\mathbf{k}, W_i) f(w)\right) = \frac{\sum_{W_i} \theta(w) f(w)}{\sum_{W_i} f(w)}.$$

Using the belief constraint (21) to substitute for the last term in the LHS, we obtain

$$\int_0^1 \min\left\{\frac{k_i}{F^-(\underline{w}_i) + z\sum\limits_{W_i} f(w')}, 1\right\} dz = \frac{\sum_{W_i} \theta(w) f(w)}{\sum_{W_i} f(w)}. \tag{22}$$

Note that $F^-(\underline{w}_i) + z\sum\limits_{W_i} f(w') \sim U[F^-(\underline{w}_i), F(\bar{w}_i)]$ with density $\frac{1}{\sum\limits_{W_i} f(w)}$ since $z \sim U[0,1]$. Hence, we can rewrite (22) as

$$\frac{1}{\sum\limits_{W_i} f(w)} \int_{F^-(\underline{w}_i)}^{F(\bar{w}_i)} \min\left\{\frac{k_i}{z}, 1\right\} dz = \frac{\sum_{W_i} \theta(w) f(w)}{\sum_{W_i} f(w)},$$

yielding condition (iv). □

Proof of Lemma 9. To prove convergence, we first partition the set of types into subsets W_i of types for sufficiently small ν as follows: (i) if we order the signal thresholds of all types, any adjacent thresholds that are within ν of each other belong to the same subset and (ii) $j > i$ implies that the thresholds associated to types in W_j are lower than those associated to W_i – by at least ν. Also, let $Q_w^\nu(\chi|\mathbf{k}^\nu, z) := \mathbb{P}\left(X \leq \chi \middle| x = k^\nu(w), \psi(X, W_i) = z\right)$ represent the beliefs about capacity of an agent of type $w \in W_i$ conditional on receiving her threshold signal $k^\nu(w)$ and on the event that the default rate in W_i is equal to z.

Note that a type-w agent receiving signal $x = k^\nu(w)$ knows that all agents with types in W_j are defaulting if $j < i$ and repaying if $j > i$. Also, the support of $Q^\nu_w(\cdot|\mathbf{k}^\nu, z)$ must lie within $[k^\nu(w) - \nu/2, k^\nu(w) + \nu/2]$. Given this, by the law of iterated expectations, her expected monitoring probability conditional on $x = k^\nu(w)$ can be written in terms of her strategic belief as follows:

$$\mathbb{E}(P|\mathbf{k}^\nu; k^\nu(w)) =$$

$$\int_0^1 \int_{k^\nu(w)-\nu/2}^{k^\nu(w)+\nu/2} \min\left\{ \frac{\chi}{F(\bar{w}) + \sum\limits_{\cup_{j<i} W_j} f(w') + z \sum\limits_{W_i} f(w')}, 1 \right\} dQ^\nu_w(\chi|\mathbf{k}^\nu, z) dA_w(z|\mathbf{k}^\nu, W_i). \tag{23}$$

In addition, notice that we can always express $\mathbb{E}(P|\mathbf{k}^\nu; k^\nu(w))$ in terms of the threshold signal $k^\nu(w)$ and relative threshold differences $\Delta_{w'} = (k^\nu(w') - k^\nu(w))/\nu$. Importantly, as Sakovics and Steiner (2012) emphasize, strategic beliefs depend on the relative distance between thresholds $\Delta_{W_i} = \{\Delta_{w'}\}_{w' \in W_i}$ rather than on their absolute distance. That is, keeping Δ_{W_i} fixed, $A_w(z|\mathbf{k}^\nu, W_i)$ does not change with ν.[24] This implies that strategic beliefs satisfy the belief constraint when $\nu = 0$.

Fix $k^\nu(w) = k_i$ for some $w \in W_i$ and fix Δ_{W_i}, for all $i = 1, \cdots, I$ and all ν sufficiently small. By fixing relative differences, the partition $\{W_i\}_{i=1}^I$ still satisfies the above definition and thus, does not change as $\nu \to 0$. We are going to show that indifference condition $\mathbb{E}(P|\mathbf{k}^\nu; k^\nu(w)) = \theta(w)$ is approximated by the limit condition in the lemma for ν sufficiently small.

Note that the inner integral in (23) is bounded below by $\min\left\{ \frac{k^\nu(w)-\nu/2}{F(\bar{w}) + \sum\limits_{\cup_{j<i} W_j} f(w') + z \sum\limits_{W_i} f(w')}, 1 \right\}$ and above by $\min\left\{ \frac{k^\nu(w)+\nu/2}{F(\bar{w}) + \sum\limits_{\cup_{j<i} W_j} f(w') + z \sum\limits_{W_i} f(w')}, 1 \right\}$. Hence,

$$\int_0^1 \min\left\{ \frac{k_i - \nu/2}{F(\bar{w}) + \sum\limits_{\cup_{j<i} W_j} f(w') + z \sum\limits_{W_i} f(w')}, 1 \right\} dA_w(z|\mathbf{k}^\nu, W_i) \leq \mathbb{E}(P|\mathbf{k}^\nu; k^\nu(w))$$

$$\leq \int_0^1 \min\left\{ \frac{k_i + \nu/2}{F(\bar{w}) + \sum\limits_{\cup_{j<i} W_j} f(w') + z \sum\limits_{W_i} f(w')}, 1 \right\} dA_w(z|\mathbf{k}^\nu, W_i). \tag{24}$$

The first term in these integrals is Lipschitz continuous. In addition, the next lemma shows that $dA_w(z|\mathbf{k}^\nu, k^\nu(w))$ is bounded for all ν.

[24] This is straightforward to check. First, if we substitute $X = k^\nu(w) - \nu\eta$ (since agents with type w get her threshold signal) and $k(w') = \nu\Delta_{w'} + k^\nu(w)$ into (15), we find that $\psi(X, W_i)$ only depends on Δ_{W_i} and $k^\nu(w)$. But this means that $A_w(z|\mathbf{k}^\nu, W_i)$ only depends on Δ_{W_i} and $k^\nu(w)$ because h is independent of ν.

Lemma 10. $0 \leq \dfrac{\partial A_w(z|\mathbf{k}^\nu, k^\nu(w))}{\partial z} \leq \dfrac{\sum_{W_i} f(w')}{f(w)}$ *for all* $w \in W_i$ *and all* z *in the support of* $A_w(\cdot|\mathbf{k}^\nu, k^\nu(w))$.

See proof below.

Hence, the LHS and the RHS of (24) uniformly converge to each other as $\nu \to 0$, leading to limit indifference conditions (20). Note also that $k^\nu(w) \in [-\bar{\nu}/2, 1+\bar{\nu}/2]$ and, keeping $\{W_i\}_{i=1}^I$ fixed, $\Delta_{w'} \in [-1,1]$ for all $w' \in W_i$. That is, the solution to the system of indifference conditions $\mathbb{E}(P|\mathbf{k}^\nu; k^\nu(w)) = \theta(w)$ lies in a compact set.[25] Accordingly, we can find $\hat{\nu}$ so that indifference conditions are within ε of the limit condition for all $\nu < \hat{\nu}$, leading to their solutions being in a neighborhood of the solution $\mathbf{k}$ of limit indifference conditions (20). □

Proof of Lemma 10. Let $\psi^{-1}(z, W_i)$ be the inverse function of $\psi(X, W_i)$ w.r.t. X. The latter function is decreasing in X as long as $0 < \psi(X, W_i) < 1$, implying that ψ^{-1} is well defined and decreasing in such a range of capacities. Since the signal of an agent of type w satisfies $x = X + \nu\eta$, we can express her strategic belief as

$$A_w(z|\mathbf{k}^\nu, W_i) = \mathbb{P}\left(\psi^{-1}(z, W_i) \leq k^\nu(w) - \nu\eta\right) = H\left(\frac{k^\nu(w) - \psi^{-1}(z, \mathbf{k}^\nu, W_i)}{\nu}\right).$$

Differentiating w.r.t. z yields

$$\begin{aligned}\frac{\partial A_w(z|\mathbf{k}^\nu, W_i)}{\partial z} &= \frac{1}{\nu} h\left(\frac{k^\nu(w) - \psi^{-1}(z, W_i)}{\nu}\right)\left(-\frac{\partial \psi^{-1}(z, W_i)}{\partial z}\right) \\ &= \frac{h\left(\frac{k^\nu(w) - \psi^{-1}(z, W_i)}{\nu}\right)}{\frac{1}{\sum_{W_i} f(w')} \sum_{W_i} H\left(\frac{k^\nu(w') - \psi^{-1}(z, W_i)}{\nu}\right) f(w')}.\end{aligned}$$

For all $z \in (0,1)$, we must have $h\left(\frac{k^\nu(w) - \psi^{-1}(z, W_i)}{\nu}\right) > 0$ because h is bounded away from zero in its support. Hence, the last term is positive and weakly lower than $\dfrac{\sum_{W_i} f(w')}{f(w)}$. □

B. Continuous distribution of returns

We now state our results under the assumptions introduced in text. Namely, we assume here that the function F is an arbitrarily fine approximation of some continuous distribution whose properties are consistent with Assumption 1. We do so by expressing equilibrium conditions from the previous section in terms of a continuous distribution F and solve them.

[25] If $\{W_i\}_{i=1}^I$ is not kept fixed then when ν is very small, $\mathbb{E}(P|\mathbf{k}^\nu; k^\nu(w))$ would be discontinuous at some ν, implying a violation of the indifference condition for some $w \in \mathcal{W}^*$.

Proof of Lemma 1. The propensity to default θ is found by equating the expected payoff from repaying to that of defaulting. From (6), the payoff from paying back the loan is $u(w,1,0) = (y+b)(w-\bar{w})$, while the expected payoff from defaulting when verification probability is P is given by $(1-P)\gamma\mu(y+b)w$. Thus, θ solves

$$(y+b)(w-\bar{w}) = (1-\theta)\gamma\mu(y+b)w,$$

which leads to $\theta = 1 - \frac{1}{\mu\gamma}\left(1-\frac{\bar{w}}{w}\right)$. □

Proof of Proposition 1. As stated in the text, uniqueness should be interpreted as the existence of a unique equilibrium in nearby discrete-F economies (Theorem 1). □

Proof of Lemma 3. The statement of the lemma is a continuous-F version of Lemma 8. □

Proof of Lemma 4. Note that the derivative of $\theta(w)F(w)$ is given by

$$\left(1 - \frac{1}{\mu\gamma}\left(1-\frac{\bar{w}}{w}\right)\right) f(w) - \frac{\bar{w}}{w^2}\frac{1}{\mu\gamma}F(w),$$

which has the same sign as

$$1 - \frac{w}{\bar{w}}(1-\mu\gamma) - \frac{F(w)}{wf(w)}.$$

If $\dfrac{F(w)}{wf(w)}$ is increasing the expression is strictly decreasing. Now, since the second term is zero at $w=0$ and $\dfrac{F(w)}{wf(w)}$ is increasing and $\lim_{w\downarrow 0}\dfrac{F(w)}{wf(w)} < 1$, the expression – and hence the slope of $\theta(w)F(w)$ – is initially positive and eventually negative for high enough w. That is, $\theta(w)F(w)$ is single peaked. □

Proof of Proposition 2. To obtain the characterization of equilibrium threshold in our model, we proceed as follows. First, we express Theorem 2 in terms of continuous type distributions. Second, we argue that the single peakedness of $\theta(w)F(w)$ implies the existence of a unique interval of types $(\bar{w}, w^*)$ such that $k(w) = \theta(w^*)F(w^*)$ for types in the interval and $k(w) = \theta(w)F(w)$ for $w > w^*$. Finally, we use the conditions in the theorem to pin down w^*. The last part of the proof simply shows that $\theta(w^*)F(w^*)$ is increasing in $\bar{w}$.

The version of Theorem 2 for continuous F implies the existence of a unique partition of types with propensity to default between 0 and 1 into intervals $\{(\underline{w}_j, \bar{w}_j]\}_{j=1}^{J}$ such that:

(a) if $k(w)$ is strictly decreasing in an interval i then it is constant in intervals $j-1$ and $j+1$ and vice versa;

(b) if $k(w)$ is strictly decreasing in interval j then $k(w) = \theta(w)F(w)$ for all $w \in (\underline{w}_j, \bar{w}_j]$;

(c) if $\theta(w)F(w)$ is not strictly decreasing in $(\underline{w}_j, \bar{w}_j]$ then $k(w) = k_j$ for all $w \in (\underline{w}_j, \bar{w}_j]$ with k_j satisfying $k_j = \theta(\bar{w}_j)F(\bar{w}_j) \geq \theta(\underline{w}_j)F(\underline{w}_j)$ (with equality if $\underline{w}_j > \bar{w}$) and

$$\int_{F(\underline{w}_1)}^{F(\bar{w}_1)} \min\left\{\frac{k_i}{z}, 1\right\} dz = \int_{\underline{w}_j}^{\bar{w}_j} \theta(w)f(w)dw. \tag{25}$$

Part (a) follows from conditions (i)-(ii) in Theorem 2, which mean that k is decreasing, so we can partition the space of types into a collection of successive intervals in which k alternates between being strictly decreasing and constant. Part (b) follows from (ii)-(iii): a strictly decreasing k in a given interval of types is approximated by a (growing) collection of consecutive, singleton W_i in the discrete economy. But then, as the mass associated to each of these singletons goes to zero, F^- approximates F, and condition (iii) implies that k converges to $\theta(w)F(w)$.

Part (c) follows from parts (a) and (b) and conditions (iii) and (iv). Since $\theta(w)F(w)$ is continuous, parts (a) and (b) imply that $k(\bar{w}) = \theta(w)F(w)$ at the boundaries of an interval in which k is constant, except possibly when $\underline{w}_i = \bar{w}$, in which case condition (iii) requires that $k_j \geq \theta(\underline{w}_j)F(\underline{w}_j)$. Expression (25) is the continuous counterpart of (iv).

We now argue that the single peakedness of $\theta(w)F(w)$ (Lemma 4 in Appendix I) leads to a partition consisting of two intervals, the first one in which k is constant and the second one in which it is strictly decreasing.

First notice that k decreasing implies that there must be at least one pooling threshold because $\theta(w)F(w)$ is initially increasing. To show why there is only one, we use the fact that condition (c) requires that $k_1 = \theta(\bar{w}_1)F(\bar{w}_1) \geq \theta(\underline{w}_1)F(\underline{w}_1)$. Given the single peakedness of $\theta(w)F(w)$ and $k(w)$ being decreasing, we must have that $\theta(w)F(w)$ is increasing at $\underline{w}_1$ and decreasing at $\bar{w}_1$. Otherwise, either $\theta(w)F(w)$ is decreasing at $\underline{w}_1$ or $\theta(w)F(w)$ is increasing in $(\underline{w}_1, \bar{w}_1)$. The former case implies that $\theta(\bar{w}_1)F(\bar{w}_1) < \theta(\underline{w}_1)F(\underline{w}_1)$, violating (c). The latter case implies that $k_1 = \max_{w\in[\underline{w}_1,\bar{w}_1]} \theta(w)F(w)$, which implies that the LHS of (25) is greater than the RHS.[26]

Accordingly, by single peakedness, if $\theta(w)F(w)$ is increasing at $\underline{w}_1$ and decreasing at $\bar{w}_1$ we cannot find another interval satisfying the same condition that does not intersect with $[\underline{w}_i, \overline{w}_i]$. Thus, there must be a unique interval of returns at which k is constant. Finally, since $\theta(w)F(w)$ is increasing in $[\bar{w}, \underline{w}_1]$, the monotonicity of $k(w)$ requires that $\underline{w}_1 = \bar{w}$.

We finish the characterization of equilibrium thresholds by showing that $\bar{w}_1 = w^*$, where w^* is the unique solution to (14) in $(\bar{w}, \infty)$ when $\bar{w} < w_{max}$.

Condition (c) implies that $k_1 \geq F(\bar{w})$. Hence, solving the integral and substituting for

[26] Since $k_1 = \max_{w\in[\underline{w}_1,\bar{w}_1]} \theta(w)F(w)$ we have that

$$\int_{F(\underline{w}_1)}^{F(\bar{w}_1)} \min\left\{\frac{k_i}{z}, 1\right\} dz > \int_{F(\underline{w}_1)}^{F(\bar{w}_1)} \min\left\{\frac{\theta(F^{-1}(z))z}{z}, 1\right\} dz = \int_{\underline{w}_j}^{\bar{w}_j} \theta(w)f(w)dw,$$

where the last equality comes from the change in variable $w = F^{-1}(z)$ ($dz = f(w)dw$).

$k_1 = \theta(w^*)F(w^*)$ and $\underline{w}_1 = \bar{w}$, we can express the LHS of (25) as

$$\int_{k_i}^{F(w^*)} \frac{k_i}{z} dz + \int_{F(\bar{w})}^{k_i} dz = k_i \log\left(\frac{F(w^*)}{k_i}\right) + k_i - F(\bar{w}) = \theta(w^*)F(w^*)(1 - \log\theta(w^*)) - F(\bar{w}).$$

Equating the RHS of the last expression to the RHS of (25) yields (14). To show that it has a unique solution in $(\bar{w}, \infty)$, we express it as

$$\theta(w^*)F(w^*)(1 - \log\theta(w^*)) - \int_{\bar{w}}^{w^*} \theta(w)f(w)dw = F(\bar{w}), \tag{26}$$

and differentiate the LHS w.r.t. w^*, which yields $(-\log\theta(w^*))\left(\theta'(w^*)F(w^*) + \theta(w^*)f(w^*)\right)$. The first term in this expression is positive, while the second term is the slope of $\theta(w^*)F(w^*)$, which is first positive then negative in $[\bar{w}, \infty)$ when $\bar{w} < w_{max}$. That is, the LHS is first increasing and then decreasing in $[\bar{w}, \infty)$. Hence, since the RHS is constant, the above expression has at most two solutions in $[\bar{w}, \infty)$. But notice that $\bar{w}$ is always a solution and that the LHS is increasing at $\bar{w}$ if $\bar{w} < w_{max}$. This, combined with the fact that the LHS approaches zero as w grows while the RHS is strictly positive, implies that there exists a unique solution in $(\bar{w}, \infty)$.

Obviously, if $\bar{w} \geq w_{max}$ then $\theta(w)F(w)$ is strictly decreasing for $w \geq \bar{w}$, and conditions (a)-(c) lead to $k(w) = \theta(w)F(w)$, i.e., to $w^* = \bar{w}$.

Finally, we need to show that $\theta(w^*)F(w^*)$ is increasing in $\bar{w}$. Note that the propensity to default θ goes up with $\bar{w}$ and that $\theta(w)F(w)$ is decreasing at w^*. Given this, if we can show that w^* goes down after an increase in $\bar{w}$ then we would have proven that $\theta(w^*)F(w^*)$ increases with $\bar{w}$. We do so by implicitly differentiating (26):

$$\frac{\partial LHS}{\partial w^*}\frac{dw^*}{d\bar{w}} = \int_{\bar{w}}^{w^*} \frac{\partial\theta(w)}{\partial\bar{w}} f(w)dw.$$

From the previous argument, we know that $\frac{\partial LHS}{\partial w^*} < 0$, while the RHS of the last expression is positive since $\frac{\partial\theta(w)}{\partial\bar{w}} > 0$. Hence, it must be that $\frac{dw^*}{d\bar{w}} < 0$. □

Proof of Lemma 5. To prove that the principal always sets $b < \infty$, first note that Proposition 2 implies the existence of a type $\hat{w} \geq \bar{w}$ such that $a = 0$ if $w < \hat{w}$ and $a = 1$ otherwise. Therefore, we can express the budget constraint (10) as follows:

$$\frac{b}{y + b} \leq \int_{\hat{w}}^{\infty} \bar{w}dF(w) + \mu\left(P + (1 - \gamma)(1 - P)\right)\int_{0}^{\hat{w}} wdF(w). \tag{27}$$

The LHS converges to one as $b \to \infty$. Hence, we need to show that the RHS is strictly less than 1 for all $\bar{w}$. Since $\hat{w} \geq w$ and the RHS is increasing in P for fixed $\hat{w}$, the RHS is bounded above by

$$\int_{\hat{w}}^{\infty} \hat{w}dF(w) + \mu\int_{0}^{\hat{w}} wdF(w), \tag{28}$$

which is strictly less than one for all $\hat{w}$ by Assumption 2. □

Proof of Proposition 3. To prove that b is increasing in X when $X \geq \theta_{\bar{w}}(w^*)F(w^*)$ first recall that $\theta_{\bar{w}}(w^*)F(w^*)$ is increasing in $\bar{w}$ by Proposition 2. Accordingly, an increase in X allows the principal to choose a higher repayment cutoff. In this context, we just need to show that higher b requires higher $\bar{w}$ to prove that b (weakly) goes up with X, since a higher X only relaxes the borrowing constraint implicit in $X \geq \theta_{\bar{w}}(w^*)F(w^*)$.[27]

To show that b goes up with $\bar{w}$, we first note that the budget constraint (27) must hold with equality. To see why, notice that the propensity to default $\theta_{\bar{w}}(\cdot)$ and hence $k(\cdot)$ do not depend on b by Proposition 1. Accordingly, for any given $\bar{w}$ the objective function (9) is strictly increasing in b since P, $\{w : a = 0\}$ and $\{w : a = 1\}$ are constant for all b, whereas payoffs under both repayment and default are strictly increasing in b. Hence, since the LHS of budget constraint (27) is strictly increasing in b, for any given $\bar{w}$, the principal chooses b so that the budget constraint binds.

In this context, two things can happen after X goes up to X' and the principal issues loan $(b', \bar{w}')$: (i) $X' \geq \theta_{\bar{w}'}(w^{**})F(w^{**})$, where w^{**} is the new upper bound of the cluster, or (ii) $X' < \theta_{\bar{w}'}(w^{**})F(w^{**})$.

To show that $b' \geq b$ when $X' \geq \theta_{\bar{w}'}(w^{**})F(w^{**})$, we need to argue that the RHS of budget constraint is increasing at the optimal $\bar{w}$ associated with X. If that is the case then a higher capacity relaxes the constraint on $\bar{w}$, in turn relaxing the constraint (27) on b and allowing the principal to give a bigger loan amount. We do so by contradiction. Assume that the RHS is strictly decreasing in $\bar{w}$ at the optimal contract associated with X. Since $X \geq \theta_{\bar{w}}(w^*)F(w^*)$ we must have that all agents with $w < \bar{w}$ default and those with $w \geq \bar{w}$ repay, implying that $P = \min\{1, X/F(\bar{w})\}$. In this context, lowering $\bar{w}$ while increasing b is feasible since the principal's revenue goes up after a reduction in $\bar{w}$ (the RHS is strictly decreasing) while the cluster threshold goes down so $X \geq \theta_{\bar{w}}(w^*)F(w^*)$ is still satisfied at the new contract. But notice that such a contract strictly increases agent expected payoffs since it increases gross returns while reducing the deadweight loss of defaults, which are reverted back to agents' payoffs given that the zero profit condition binds.[28] Accordingly, b must be increasing in $\bar{w}$ at the optimal contract associated with X.

Finally, if $X' < \theta_{\bar{w}'}(w^{**})F(w^{**})$, it must be that $\bar{w}' > \bar{w}$ because $\theta_{\bar{w}'}(w^{**})F(w^{**}) > \theta_{\bar{w}}(w^*)F(w^*)$. But then, by the same argument as above, it must be that $b' > b$, otherwise, agents' payoffs would go down with respect to the contract $(b, \bar{w})$, which is still feasible. □

[27] It can easily be shown that b strictly goes up when the constraint strictly binds if the optimal payoff to entrepreneurs as a function of $\bar{w}$ are quasiconcave.

[28] Formally, agents' payoffs are given by $\int_{\bar{w}}^{\infty}(y+b)(w-\bar{w})dF + P\gamma\mu\int_0^{\bar{w}}(y+b)wdF$. Since the budget constraint holds with equality, we can express the last term as

$$P\gamma\mu\int_0^{\bar{w}}(y+b)wdF = b - \int_{\bar{w}}^{\infty}(y+b)\bar{w}dF - (1-\gamma)\mu\int_0^{\bar{w}}(y+b)wdF.$$

Hence, agents' payoffs can be written as

$$\int_{\bar{w}}^{\infty}(y+b)wdF + b - (1-\gamma)\mu\int_0^{\bar{w}}(y+b)wdF,$$

which strictly go up if we increase b and lower $\bar{w}$.

Proof of Proposition 4 in Appendix I. First consider the case of $X < \underline{X}$. If $\bar{w} < w_{max}$, it must be that (14) has only solutions, $w_1 < \bar{w}$ and $w_2 > w_{max}$. The former cannot be an equilibrium since it would require $\psi = F(w_1) < F(\bar{w})$, i.e., that some agents that strictly prefer to default choose to repay, and hence equilibrium is unique. The same argument applies when $\bar{w} \geq w_{max}$. If $X = \underline{X}$ then $w_1 = \bar{w}$ is also an equilibrium.

Second, let $X > \bar{X}$. In this case, $\theta_{\bar{w}}(w)F(w)$ lies below X, implying that for any given P such that all agents with default propensity less than P default, there is enough capacity so that the monitoring probability is higher than P. Thus, equilibrium is unique and involves $\psi = F(\bar{w})$. If $X = \bar{X}$ then w_{max} is a solution of (14), representing a second equilibrium.

Finally, if $X \in (\underline{X}, \bar{X})$ there are three equilibria given by the two solutions in $[\bar{w}, \infty)$ to (14) and another equilibrium with $\psi = F(\bar{w})$ since $F(\bar{w}) < X$, and thus, the principal can credibly sustain $P = 1$ in equilibrium. □