Financial Contracting with Enforcement Externalities

Lukasz A. Drozd and Ricardo Serrano-Padial†

ABSTRACT

Enforcement of financial contracts often depends on the collective behavior of agents, leading to a feedback loop that can propagate economic shocks and/or interact with other frictions. For example, aggregate default rate may influence the speed of legal enforcement, or liquidity of the collateral, affecting incentives to honor financial contracts. Here we extend costly state verification framework to think about a friction that links enforcement to collective behavior of agents. The key element of our model is enforcement in the form of accumulated capacity. Our analysis sheds new light on such phenomena as credit crunches and the link between accumulation of enforcement infrastructure, economic growth and political economy frictions.

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, Pompeu Fabra, Philadelphia Fed, Queen Mary, Royal Holloway, St. Louis Fed, Wharton, Surrey, and UW-Madison for their comments. Drozd acknowledges the financial support of the Rodney White Center at the Wharton School of the University of Pennsylvania. 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.
1. Introduction

Enforcement of financial contracts often depends on the collective behavior of an entire population of agents, leading to a feedback loop that can propagate economic shocks and/or interact with other frictions. The recent foreclosure crisis illustrates the relevance of this issue. As is well known, during the crisis rising foreclosure rate made it increasingly difficult for lenders to timely foreclose on delinquent borrowers, delaying the threat of eviction by as much as 3 years in troubled areas.\textsuperscript{1} The implicit transfer associated with foreclosure delays directly incentified defaulting, which available evidence suggests raised the national foreclosure rate by as much as 25\% or more.\textsuperscript{2}

The foreclosure crisis is hardly an isolated episode. In fact, most financial crises test the limits of the economy’s enforcement infrastructure.\textsuperscript{3} Not surprisingly, the idea of propping up the economy’s resolution mechanisms has always been the cornerstone of policy response to financial crises. In the face of an overwhelming empirical evidence on the link between the flow of credit in the economy and the efficacy of contract enforcement, our reading of this evidence

\textsuperscript{1}The national average delay was about about 9 months. For detailed discussion of foreclosure delays, see Cordell et al. (2013); Herkenhoff and Ohanian (2012). However, the national average likely masks the severity of this problem as in troubled areas the delays were much longer. For example, in the midst of the crisis it took almost three years (1,027 days) to foreclose in Florida. In D.C., foreclosure averages 1,053 days and delinquent borrowers in New York often stay in their homes for an average of 906 days. See “ Delaying foreclosure: Borrowers keep homes without paying”, CNN Money, Dec. 28, 2011, \url{http://money.cnn.com/2011/12/28/real_estate/foreclosure/}.

\textsuperscript{2}Using a quantitative model, Herkenhoff and Ohanian (2012) recently estimate that ensuing delays added as much as 25\% to the aggregate delinquency rate during the crisis. In a related empirical study, Chan et al. (2015) show that in the subprime pool of loans, a delay of 9 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. For a thorough discussion of foreclosure delays see Cordell et al. (2013).

\textsuperscript{3}For instance, during the Asian crisis the volume of nonperforming loans reached 14, 27, 32 and 51 percent of GDP for Korea, Indonesia, Malaysia, and Tailand, respectively (Woo, 2000). Liquidation of these assets took years and no significant improvement in credit conditions occurred before this was resolved (Mark, 2000; Enoch et al., 1998). In the case of the 1995 Mexican crisis, Krueger and Tornell (1999) argue that delayed resolution was the chief among reasons responsible for the slow recovery of the Mexican economy.
is that a careful exploration of this friction is needed.\textsuperscript{4}

To that end, our paper considers a simple friction that endogenously links ex post effectiveness of contract enforcement to the collective behavior of agents. The key building block of our new theory is the idea of costly state verification (Townsend, 1979)—the workhorse model of endogenously incomplete markets and financial frictions in much of the economics literature. However, unlike in the original framework, enforcement takes the form of ex ante accumulated capacity, and thus is subject to depletion in the short-run. If that’s the case, agents become aware that any marginal increase in the aggregate default rate weakens enforcement, and behave strategically. We refer to this central feature of our model as the \textit{enforcement externality} and study its implications for optimal debt contracting.

Our friction captures the basic notion that disrupted or delayed enforcement generally benefits borrowers. In practice, such a disruption may be caused by the fact that legal capacity to assist private markets in enforcing contracts is a scarce resource. In the context of individual markets, it can be caused by the fact that the value and liquidity of collateral crucially depends on the amount of collateral that is being liquidated. The value and liquidity of collateral is broadly important for enforcement, as it influences incentives to obey contracted terms. In practice, both channels may operate in tandem.

The problem we highlight here is so far not very well understood in economics literature. The papers that consider optimal debt contracting, such as the original contributions due to

\textsuperscript{4}The link between enforcement and access to credit is a well established empirical fact. For example, Safavian and Sharma (2007) and Jappelli et al. (2005) use firm level 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 time constraints in busy bankruptcy courts. In a related paper, Ponticelli (2015) shows that court congestion changes the impact of financial reforms increasing lender protection on firms’ access to credit. For cross-country evidence, see Djankov et al. (2007).
Townsend (1979), Gale and Hellwig (1985)—while studying a setup closely related to ours—all assume constant enforcement cost per contractual relation. While the financial accelerator model by Bernanke et al. (1999) involves an externality, the nature of the externality is fundamentally different from ours. Our paper also contrasts with much of the literature which studies the effect of enforcement capacity constraints, which thus far has only focused on stylized environments featuring homogeneous agents. Since agent heterogeneity naturally interferes with the coordination problem that such friction gives rise to, our paper extends the domain of applicability of existing analysis to a much broader spectrum of economic environments featuring virtually unrestricted form of agent heterogeneity. To name a few notable contributions to this closely related literature that we build on, propagative effect of enforcement capacity constraints has been studied in the context of micro-finance literature by Bond and Rai (2009), spillovers of private default to sovereign default by ?, economics of crime by (Bond and Hagerty, 2010), and in the context of tax evasion by (Bassetto and Phelan, 2008) and the literature this paper initiated.5

Formally, our model involves a principal-multiagent environment, and is most closely related to setup in Gale and Hellwig (1985).6 Ex ante, the principal chooses enforcement capacity. Ex post, the principal learns the aggregate shock on pre-existing loans, and this shock exogenously reduces the capacity that can be later used to enforce new loans. Having this information, the principal makes optimal loans to a population of ex ante homogeneous entrepreneurs.

5The most closely related paper is Bond and Rai (2009). The paper explores a microfinance model of borrower runs. The run is generated by the fact that when more agents choose to default on a given lender, they exhaust her resources and take away the ability of this lender to sustain the promised continuation value. This, in turn, incentivizes defaulting.

6An important difference between our environment and Gale and Hellwig (1985) is that our model involves stochastic monitoring and a different notion of the role of monitoring in the economy (divorced from the idea of liquidation).
Entrepreneurs then privately observe their *heterogeneous* output, and simultaneously decide whether to repay the loan or not. Contracts are enforced at the last stage, and in case enforcement capacity is binding, the decision to default is subject to aforementioned *enforcement externality*. That is, when the constraint binds, the decision problem is subject to strategic complementarity. Finally, default is socially wasteful, and the principal would like to minimize it.

The effect of the enforcement externality we analyze here if far from being well understood in the economic literature. Most quantitative models of financial frictions abstract from such a friction by assuming that costs of defaulting and enforcement are scale invariant.\(^7\) On the other hand, the literature that does explore the effect of constrained enforcement of financial contracts focuses on models featuring homogeneous agents, which limits the insight to a qualitative one because agent homogeneity leads to extreme outcomes in which either everyone or no one defaults.\(^8\) In contrast, our theory features a population of agents who differ in terms of their (ex post) propensity to default, as is the case in most models of credit markets. This is arguably a central element of our analysis since it captures the relationship between heterogeneity and the strength of the enforcement externality.

From a methodological perspective, we overcome here two challenges that the analysis of

\(^7\)For example, this is the baseline assumption of the financial accelerator model (Bernanke et al., 1999), widely used to study credit crunches.

\(^8\)In the finance and macroeconomic literature, two contributions stand out: Bond and Rai (2009) and Arellano and Kocherlakota (2014). Bond and Rai (2009) explore a microfinance model of borrower runs. The run is generated by the fact that when more agents choose to default on a given lender, they exhaust her resources and take away the ability of this lender to sustain the promised continuation value. This, in turn, incentivizes defaulting. Arellano and Kocherlakota (2014) study how private sector defaults may cause sovereign default. Central to their analysis is a notion of multiplicity of equilibria, which arises precisely because enforcement ability is endogenously constrained. In addition, the literature on tax evasion explores the role of limited audit capacity on optimal taxation, e.g., Bassetto and Phelan (2008).
such a friction involves. First, as is well known, strategic complementarities introduce multiple equilibria when the principal’s enforcement capacity is common knowledge. Multiplicity of equilibria prevents the analysis how the presence of such a friction may endogenously propagate shocks to economic fundamentals. To address 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 based on the notion of (infinitesimal) miscoordination risk. Second, heterogeneity complicates the characterization of the unique equilibrium. To address this problem, we build on the work of Sakovics and Steiner (2012) on games with symmetric equilibria and generalize their insight to games where heterogeneity naturally leads to asymmetric equilibria. This approach allows us to prove our main theoretical result: the characterization of equilibrium default strategies as a function of credit levels and enforcement capacity.

Substantively, despite heterogeneity in default propensities equilibrium surprisingly exhibits a great deal of economic fragility. Specifically, we show that for typical return distributions a significant mass 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.

Intuitively, the clustering of default decisions is generated by a domino effect driven by agents’ divergent beliefs about enforcement levels. Specifically, given the same information

---

9The model of Sakovics and Steiner (2012) allows for heterogeneous preferences but imposes preference restrictions to ensure that equilibria are symmetric. In the context of our model, this would imply that either all agents or none default. In contrast, equilibrium in our model typically involves some agents repaying and some defaulting.
about capacity, agents who are intrinsically less prone to default, i.e., those with high project returns, expect a higher default rate than more prone-to-default agents. This is because low propensity types anticipate that whenever they find default optimal so do higher propensity types, implying that lower propensity types expect lower enforcement levels. If these belief differences are big enough they induce lower propensity types to default at the same time as the higher propensity agents and a default cluster arises. Belief differences increase with the relative presence of higher propensity types. 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 can trigger a wave of coordinated defaults. The second implication of our model is that, due to the dead-weight 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 credit constraints of new borrowers as to reduce their intrinsic propensity to default and prevent a default wave. This prediction is relevant for monetary policy, as it implies that lowering interest rates need not be the most effective tool in this case. Instead, unconventional policies might be more effective. Finally, clustering lowers the effect of capacity increases on credit and thus raises the opportunity cost of investments in capacity. Consequently,
enforcement externalities can have a strong amplification effect of political economy distortions on economic development by disincentivicing the buildup of an enforcement infrastructure 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, as discussed in 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 recent financial crisis, 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 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 she 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.\(^\text{10}\)

The second application illustrates the amplification effect that enforcement externalities have on other frictions. Specifically, we look at the impact of political economy distortions on the efforts of developing economies at building enforcement infrastructure. This application is inspired by the recent workhorse model of political economy and development (Besley and Persson, 2010, 2009, 2011), which is centered around 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

\(^{10}\)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 important, as it addresses the criticism of financial accelerator models raised by Carlstrom et al. (forthcoming).
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 a lack of investment in capacity and virtually no access to credit.

Finally, our model provides additional policy recommendations that are worth emphasizing. During periods of financial distress debt sustainability is a concern due to 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.

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 agents take from the principal and $\bar{b}$ is the amount agents are ex post required to repay to the principal. It is assumed that in order to apply for a loan agents have their own equity equal to $y$ already invested in the proposed 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) \(\frac{F(w)}{wf(w)}\) is increasing and \(\lim_{w \downarrow 0} \frac{F(w)}{wf(w)} < 1\); 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 below our approach to equilibrium characterization applies to any \(F\).

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

The principal captures the full liquidation only when she monitors the project and verifies the realization of \(w\). Specifically, when liquidation is monitored by the principal, it is assumed that the entire liquidation value of the project accrues to the principal. 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 \(\gamma\) 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 \(\gamma\) leads to stronger enforcement externalities since it increases the benefits of unmonitored defaults. The

\(^{11}\)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.
principal determines her monitoring capability by accumulating enforcement capacity $X_{−1}$ in the first period, which determines the mass of defaulting agents she can monitor later on.

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

**Assumption 2.** $\int_{w'}^{\infty} w'dF(w) + \mu \int_0^{w'} wdF(w) < 1$ for all $w' \geq 0$.

We next lay out the timing of events and specify actions and payoffs of all agents.

**A. Timing, constraints, and payoffs**

1. **Capacity accumulation:** At the accumulation stage, the principal chooses how to allocate her exogenous endowment of ex ante resources $R > 0$. These resources can be allocated to principal’s own consumption or used to build economy’s enforcement capacity $X_{−1}$. Enforcement capacity is accumulated at a cost $c(X_{−1})$, where $c(\cdot)$ is convex, continuous and increasing. Formally, the principal chooses enforcement capacity $X_{−1} \geq 0$ and her consumption $g \geq 0$ subject to

$$R = c(X_{−1}) + g. \quad (1)$$

After this decision has been made, an aggregate shock is observed. This shock depletes a random amount $s > 0$ of principal’s capacity $X_{−1}$, and in a crude way captures the idea that the principal has some pre-existing loans on her books, and when these loans default more frequently, less capacity is available to support new loans. The residual capacity is $X = X_{−1} − s$.

In this *baseline timing* the shock occurs before the contracting stage. However, we discuss
later what happens under 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 their 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 \((b, \bar{w})\).\(^{12}\)

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 \(\mu(y+b)\). 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, agent’s utility function depends on the agent’s decision to repay \(a \in \{0, 1\}\)

\(^{12}\)We do not analyze here the optimality of debt contracts, but indeed debt contracts are optimal in the sense of non-commitment setup by Krasa and Villamil (2000). The proof is available upon request. One important caveat applies: it must be assumed that the principal cannot sign contracts that discriminate agents on the basis of non-essential characteristic of the agent (name, color of skin, etc...). Such specification would unravel our externality. See the more detailed discussion in footnote 13.
\( (a = 1 \text{ means repayment}) \) and on monitoring \( m \in \{0, 1\} \) \((m = 1 \text{ means monitoring})\):

\[
u(a, w, m) :=
\begin{cases}
(y + b)(w - \bar{w}) & a = 1 \\
\gamma \mu (y + b) w & a = 0 \& m = 0 \\
0 & a = 0 \& m = 1.
\end{cases}
\]  

(2)

After the repayment decision is made by the agents, the principal allocates her enforcement capacity to monitor agents who did not repay. The default rate \( \psi \) is given by the mass of agents who default:

\[
\psi := \int_{\{w : a(w) = 0\}} f(w) dw.
\]

(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.\(^{13}\) Accordingly, the probability that a liquidated project is monitored is given by

\[
P := \min \left\{ \frac{X}{\psi}, 1 \right\}.
\]

(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 \( \psi \). 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

\(^{13}\)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 non-discrimination laws and 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 (2013).
levels and making default more attractive.

Having described the timing and payoffs, we now turn to the description of the underlying optimization problems. We start from the agent’s and then discuss the principal’s problem. All proofs are relegated to the Appendix.

B. Agent’s problem

Since agents are homogeneous at the contracting stage, the principal offers the same loan \((b, \bar{w})\) to all agents. We do not impose any individual rationality constraints, given that the principal is benevolent and can always offer a null contract. Hence, the only relevant decision that each agent makes in our model is whether to repay the loan or not after observing her project return. An agent’s repayment decision \(a\) solves

\[
\max_{a \in \{0, 1\}} \{P^e u(w, a, 1) + (1 - P^e) u(w, a, 0)\},
\]

where \(P^e\) denotes the expected monitoring probability of the agent.

The above maximization problem implies that agents repay their loans if \(P^e\) exceeds a well defined cutoff value. The cutoff that the agent uses is denoted by \(\theta_{\bar{w}}(w)\). We refer to as the \textit{propensity to default} of an agent of type \(w\).
Lemma 1. The default decision an agent with contract \((b, \bar{w})\) is\(^{14}\)

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

\hspace{100pt} (6)

where

\[
\theta_{\bar{w}}(w) := 1 - \frac{1}{\mu \gamma} \left( 1 - \frac{\bar{w}}{w} \right). 
\]  

\hspace{100pt} (7)

Observe that the propensity to default \(\theta_{\bar{w}}(w)\) 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, i.e. \(P^e = P\), the aggregate default rate \(\psi\) defined in equation (3) is equal to \(F(\hat{w})\), where \(\hat{w}\) solves \(P = \theta_{\hat{w}}(\hat{w})\). However, when agents receive a noisy signal about \(X\), as it is going to be the case in our global game version of the model, \(P^e\) differs across agents so understanding how these expectations are formed will be key to solve 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 since their incentives to default depend on their beliefs about the behavior of other agents, i.e., about the default rate. Accordingly, 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 \((\hat{w}, \bar{w}/(1 - \gamma \mu))\). Agent types outside\(^{14}\)

\(^{14}\)Given the continuity of \(F\) we can assume without loss that an indifferent agent always chooses to repay.
this range either never default \((w \geq \bar{w}/(1 - \gamma\mu))\) or always default \((w \leq \bar{w})\).

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.]

C. Principal’s problem

The principal takes the behavior of agents as given and maximizes their 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 towards her consumption and away from accumulation of enforcement capacity.\(^{15}\)

The principal’s optimization problem is sequential. In the first period, the principal allocates ex ante resources \(R\) between \(g\) and \(X_{-1}\). 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 thus effectively contingent on \(X\).

The payoff from these two decisions is connected by the preference parameter \(\alpha\). Hence, it could be thought of as a choice problem of two entities who may or may not fully internalize the problem of the other entity. For example, the first stage may describe the choice of a government who builds up an economy-wide enforcement infrastructure, while the second stage describes a competitive lending industry that takes the enforcement infrastructure as given. An alternative

\(^{15}\)Conversely, \(\alpha < 1\) can be interpreted as a bias in favor of promoting private investment at the expense of alternative uses of public resources.
interpretation is that both decisions are made by the lending industry, but due to some kind of market imperfection not all the trade-offs of accumulating capacity are internalized ($\alpha > 1$).

Formally, in the first period the principal chooses $X_{-1}$ and $g$ to solve

$$\max_{X_{-1},g} [\alpha g + \mathbb{E}\Pi_s(X_{-1} - s)]$$

subject to $R = c(X_{-1}) + g$ and $g \geq 0$, where $\Pi_s$ denotes the net expected utility of entrepreneurs conditional on shock $s$. In the second period the principal chooses contract $(b, \bar{w})$, given $X = X_{-1} - s$, to maximize the utility of entrepreneurs, i.e.,

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

subject to resource feasibility

$$b \leq \int_{\{w:a(w)=1\}} (y + b)\bar{w}dF +$$

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

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

In the above problem, observe that in the second period the principal maximizes agents’ net expected payoff. That is, even if agents behaves opportunistically and default strategically, their utility is fully internalized by the principal. It is the deadweight loss from defaulting what 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 solely driven by \( \bar{w} \) as in standard models of costly state verification. It is when \( \gamma \) is sufficiently high so that providing a significant amount of credit requires monitoring to prevent the deadweight loss associated to 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 \( \gamma \). Consequently, building enforcement capacity is tied in our model to the presence of enforcement externalities.\(^{16}\)

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 given the one above, 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 (\( \alpha \)) in Section 4.

\(^{16}\)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 \( X > 0 \).
A. 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. We discuss the implications of the enforcement game for the ex ante contracting problem and the choice of enforcement capacity in the next section.

It is not difficult to see that under common knowledge of $X$ the presence of enforcement capacity constraint (4) can lead to multiplicity of equilibria for a wide range of $X$. Intuitively, if agents anticipate a default rate $\psi$ higher than $X$ they expect (4) to bind, which can lead to a monitoring probability $P$ low enough to incentivize default by a mass $\psi$ of agents. Similarly, a 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 $\eta$ 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 $[0, 1]$.$^{17}$ Loosely speaking, the noise represents an agent’s “uncertainty” or lack of perfect confidence in the inference they may be making about fundamentals, such as $X$ or the loan contract.

Our first result establishes that equilibrium in this setup is unique, and that it takes a simple form of a type-identical threshold strategy $k(w)$ defined on the signal space. This implies that, if the signal exceeds the type specific cutoff level $k(w)$, all agents of this type choose to repay, and if not, they default.

**Proposition 1.** The enforcement game has a unique equilibrium.$^{18}$ Equilibrium strategies are characterized by a cutoff $k(w)$ defined on the signal space, such that:

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

The proof of the above result is standard and follows closely the one in Frankel, Morris and Pauzner (2003). We thus do not devote too much time to it. The basic idea is that agents who receive a very high signal or a very low signal follow a dominant strategy. The upper and lower dominance regions give rise to a successive elimination of dominated strategies that converges to a single common threshold from above and below.$^{19}$

While the above result goes a long way in characterizing equilibrium strategy profiles, to characterize equilibrium we must characterize $k(w)$ as function of agent type $w$. To do so, we

---

$^{17}$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.

$^{18}$Equilibrium strategies are unique up to sets of measure zero.

$^{19}$For more detailed intuition see Carlsson and van Damme (1993). Our proof in the appendix, while it relies on the same basic insight, is not particularly illuminative.
must characterize the indifference condition which states that an agent who receives a signal equal to her cutoff (threshold signal thereafter) must be indifferent between defaulting or not,

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

This, however, turns out to be a challenging task because. The issue is that this condition depends on the unknown distribution of \( P \). Since \( P \leq X/\psi \), this object involves the expectation of \( X/\psi \) by an agent who receives a signal \( x = k(w) \). The information about \( X \) is arbitrarily precise, since \( \nu \to 0 \), but determining the expectation of \( \psi \) involves strategic beliefs about the behavior of others, which in turn is a function of how their signals relate to their own thresholds \( k(w) \).

To see that it is indeed heterogeneity that makes this condition not tractable, compare what would happen if agents were homogenous and used a single threshold \( k(w) = k \)—which in such a case follows form the proposition above.

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 \( x = X + \nu \eta \) is equal to the cutoff \( k \) believes that the strategic default rate is given by \( H(\eta) \) regardless of \( \nu \), i.e., by the mass of agents with signal noise lower than hers. But since she does not observe \( \eta \), she views \( H(\eta) \) as a random variable, which is distributed uniformly in \([0, 1]\) (recall, the cdf of any random variable is uniformly distributed. This feature of beliefs is known as the Laplacian property (Morris and Shin, 2003).

Unfortunately, whereas heterogeneity does not affect uniqueness, Laplacian property no
longer holds. The reason is that agents with different types may use different thresholds and hence the distribution of the strategic default rate conditional on \( x = k(w) \) is no longer uniform. This is not necessarily a problem when solving for \( E(P|x = k(w)) \) when \( k(w) \) is not within \( \nu \) 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 \( \nu \) of each other i.e., those whose thresholds converge to the same limit threshold as \( \nu \downarrow 0 \): we do not know when such a clustering may occur without solving for (17), and to solve these conditions we must know the probability distribution underlying the expectation operator, which can be quite complex.

What allows us to nonetheless characterize the equilibrium is a powerful result established in a context of a much simpler model by Sakovics and Steiner (2012). Sakovics and Steiner (2012) showed that which the Laplacian property does not apply to any individual type, it still applies on average, as we explain below. We use this remarkable result to circumvent the need to pin down individual beliefs and fully 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. Specifically, we use the joint average condition

\[
\int_{w \in W'} E(P|x = k(w))f(w)dw = \int_{w \in W'} \theta_{\bar{w}}(w)f(w)dw, \tag{12}
\]

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 \( \psi \) by the average belief in the cluster. Since \( X = k(w) \) in the
limit when \( x = k(w) \), we can analytically characterize the average expectation in (18) and pin down \( k(w) \).

Specifically, Sakovics and Steiner (2012) establish that if we average the beliefs of agents when they receive their threshold signal \( x = k(w) \), weighted by their presence in the population \( (f(w)) \), the distribution is uniform. In other words, if we randomly select a sample of agents and ask them for their beliefs about the distribution of \( \psi \) when they receive their threshold signal, the average answer is uniformly distributed. 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. While 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 player 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.

\[
\psi(W', X) := \frac{1}{\int_{W'} f(w)dw} \int_{W'} H \left( \frac{k(w) - X}{\nu} \right) f(w)dw.
\]

Then, for any \( z \in [0, 1] \),

\[
\frac{1}{\int_{W'} f(w)dw} \int_{W'} \mathbb{P}_w (\psi(W', X) \leq z | x = k(w)) f(w)dw = z,
\]

(13)
where $P_w (\cdot | x = k(w))$ is the probability assessment of $\psi(W', X)$ by an agent receiving $x = k(w)$.

The belief constraint is instrumental to characterize equilibrium thresholds in our model as $\nu$ goes to zero. As highlighted above, it allows us to express the average indifference condition (18) of types in cluster $W'$ in a closed form and identify any clusters that may arise in equilibrium. As Proposition 5 states below, the resulting equilibrium threshold is closely related to the shape of $\theta_w(w)F(w)$. This function represents the level of enforcement capacity $X$ that makes agents of type $w$ indifferent between defaulting or not when $X$ is common knowledge and all agents with returns lower than theirs default ($\psi = F(w)$).\footnote{Under common knowledge of $X$, $\psi$ is common knowledge in equilibrium. Hence, when $\psi = F(w)$ indifference implies $P^e = X/F(w) = \theta_w(w)$.} In this regard, Assumption 1 leads to the existence of a unique cluster by ensuring that $\theta_w(w)F(w)$ 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 $\theta_w(w)F(w)$ is increasing at 0 and single-peaked.

Let $w_{\text{max}} = \arg\max_w \theta_w(w)F(w)$.

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

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

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

$$\theta_w(w^*)F(w^*) (1 - \log \theta_w(w^*)) - F(\bar{w}) = \int_{\bar{w}}^{w^*} \theta_w(w)f(w)dw. \quad (14)$$
Furthermore, $\theta_{\bar{w}}(w^*)F(w^*)$ is increasing in $\bar{w}$.

![Equilibrium Strategy in the Global Game](image)

**Figure 1**: Equilibrium Strategy in the Global Game.

Figure 3 illustrates the equilibrium strategy implied by Proposition 5. 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, this is not the case on the interval $w \in [\bar{w}, w^*]$, where agents from the interval cluster on 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 a cluster of agents with types between $w$ and $w^*$ simultaneously decide to strategically default.

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-θ_\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 θ_\bar{w}, inducing them to default too. 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 weak 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 θ_\bar{w}(w)F(w) being single-peaked, which is 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) = θ_\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 θ_\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.21

A remaining question is that, if all agents in the cluster use the same threshold, how is it possible to simultaneously satisfy their heterogeneous indifference conditions? The answer is that, away from the limit (ν > 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 5 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

\[21\text{Multiple peaks due to a multimodal distribution of returns or to non-monotonic default propensities can lead to several clusters. Theorem 2 shows how to characterize them.}\]
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 \((\hat{w}, \hat{w}/(1 - \gamma \mu))\) by earlier Lemma 2.

What allows us overcome this difficult is by Sakovics and Steiner (2012) who showed that when heterogeneous agents use different strategies that converge to the same strategy profile (same threshold \(k\)) in the limit (as signal noise \(\nu\) is taken zero), their belief about behavior of other agents obeys the Laplacian property on average rather than individually. That is, while beliefs of any individual agent who receives a threshold signal may be arbitrarily complex, on average they average out across agent types and so the Laplacian property applies on average to the group as a whole. We generalize this remarkable result here by additionally highlighting that its applicability extends far beyond the particular setup of a symmetric equilibrium analyzed in Sakovics and Steiner (2012). In particular, it can be applied to environment that yield asymmetric equilibria, as it is not necessary to know a priori which agent types use thresholds that fall under the umbrella of the signal noise.

**Lemma 5** (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.

\[
\psi(W', X) := \frac{1}{\int_{W'} f(w)dw} \int_{W'} H \left( \frac{k(w) - X}{\nu} \right) f(w)dw.
\]

Then, for any \(z \in [0, 1]\),

\[
\frac{1}{\int_{W'} f(w)dw} \int_{W'} \mathbb{P}_w (\psi(W', X) \leq z | x = k(w)) f(w)dw = z,
\]

(15)
where $P_w (\cdot | x = k(w))$ is the probability assessment of $\psi(W', X)$ by an agent receiving $x = k(w)$. That is, the Laplacian property holds on average within this arbitrary set $W'$.

Equipped with the above lemma, we are now able to fully characterize the equilibrium strategy profile $k(w)$, which is the main result of our paper. The belief constraint allows us to replace (17) by an integrated out indifference conditions given by $\int_{w \in W'} E(P^e|x = k(w)) f(w) dw = \int_{w \in W'} \theta_w(w) f(w) dw$, where $W'$ is an unknown subset of agent types who use the same threshold $k(w)$ as $\nu \downarrow 0$. According to the lemma the expectation operator of this group as a whole is uniform. Since, if they fall under same signal noise they converge to the same threshold, we can use an integrated out indifference condition without loss of generality. This is because all agents in $W'$ by definition converge to the same strategy threshold.

**Proposition 3.** There exists $w^* \geq \bar{w}$ such that

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

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

$$
\theta_w(w^*) F(w^*) (1 - \log \theta_w(w^*)) - F(\bar{w}) = \int_{\bar{w}}^{w^*} \theta_w(w) f(w) dw. \quad (16)
$$

Furthermore, $\theta_w(w^*) F(w^*)$ is single peaked by Assumption 1 and $\theta_w(w^*) F(w^*)$ is increasing in $\bar{w}$. (Note here that all agents with $w < \bar{w}$ default, regardless of the action of other agents in the economy.)
Figure 2: Equilibrium Strategy in the Global Game.

The equilibrium strategy features a bang-bang property. That is, for sufficiently high $X$, the equilibrium involves low default rate that is Pareto efficient. However, for $X$ lower than $\theta_{w^*}F(w^*)$, the default rate discontinuously jumps because a whole segment of agent types between $w$ and $w^*$ simultaneously decides to default. This result arises because agent types between $w$ and $w^*$ “cluster” on the same strategy profile, despite the fact that they are widely heterogenous in terms of their intrinsic propensity to default $\theta_{w^*}$. The default rate associated with this point is inefficient, as it is propelled by the externality.

Figure 3 illustrates the equilibrium strategy implied by Proposition 5. It plots $k(w)$ in relation to $\theta_{w}(w)F(w)$. By assumption 1, we know that $\theta_{w}(w)F(w)$ is a single-peaked function as illustrated.

As is clear from the figure, $k(w)$ traces $\theta_{w}(w)F(w)$ for all $w > w^*$. However, this is not the case on the interval $w \in [\bar{w}, w^*]$, where agents from the interval ‘cluster’ on the same threshold $k(w) = k(w^*)$. 

28
The reason why agents tend to cluster on the same strategy profile is as follows. As the signal noise $\nu$ vanishes, agents with high propensity $\theta_{\bar{w}}$ who use a smaller signal threshold $k(w)$ than agents with slightly lower $\theta_{\bar{w}}$, would already be certain that all higher propensity types are defaulting when they receive their threshold signal. If these high-$\theta_{\bar{w}}$ types have a sufficiently strong presence in the population, the implied by this fact increase in the perceived aggregate default rate by the low-$\theta_{\bar{w}}$ might be enough to push their respective expected monitoring probability below their threshold $k(w)$, inducing them to default and contradicting that their respective threshold is above the one of the high-$\theta_{\bar{w}}$ type. This effect “snowballs” down the distribution of propensity types until the subsequent types a sufficiently weak presence in the population so that higher default rate no longer compensates for their lower propensity to default.

Formally, the condition for clustering is that $\theta_{\bar{w}}(w)F(w)$ is single-peaked. To see why, note that if cutoffs are strictly decreasing for all $w$, an agent who receives her threshold signal is certain that the default rate is $F(w)$. Hence, she defaults if and only if $k(w) = \theta_{\bar{w}}(w)F(w)$. But notice that single-peakedness is driven by the fact that $F$ initially goes up fast enough with $w$ to more than compensate for the decrease in propensity to default $\theta_{\bar{w}}(w)$, so that the product of the two increases towards the peak. Hence, single-peakedness of $\theta_{\bar{w}}(w)F(w)$ and the described above snowballing effect are invariably related.

A remaining question is that, if all agents in the cluster use the same threshold, how is it possible to simultaneously satisfy their heterogeneous indifference conditions? The answer is quite simple. Away from the limit ($\nu > 0$), thresholds of types in the cluster are actually heterogeneous within $\nu$ of each other so that different types have different expected monitoring
probabilities conditional on receiving their threshold signals. However, in the limit, none of they converge to the same threshold strategy.

Our first result shows that there is a unique equilibrium, given by threshold strategies.

**Proposition 4.** The enforcement game has a unique equilibrium at any $\nu$ sufficiently close to zero.\(^\text{22}\) 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 rate, hindering agents’ ability to sustain multiple equilibria by seamlessly coordinating their beliefs about $\psi$.\(^\text{23}\) The proof is fairly standard, and roughly follows the logic in Frankel, Morris and Pauzner (2003). To provide some intuition let us focus on equilibria in which agents follow threshold strategies. Equilibrium thresholds must satisfy indifference conditions

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

that is, when receiving her signal cutoff $x = k(w)$, an agent is indifferent between defaulting or not. Now, consider first the symmetric case in which all strategic agents use the same signal

\(^{22}\)Equilibrium strategies are unique up to sets of measure zero.

\(^{23}\)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 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 $\nu$ away from her own signal, and thus that they admit a possibility that $X$ is as far as $3\nu$ away from her signal, thus placing positive probability that other agents think other agents observe signals as far as $2\nu$ away from her signal. When this reasoning is repeated infinitely many times, it is clear that infinite order beliefs about $X$ will fan out for any arbitrarily small $\nu$. This divergence of higher order beliefs translates into divergent beliefs about $\psi$.  

30
cutoff $k(w) = k$. In this simple case the default rate of strategic agents (those with $\theta_{w}$ 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 $x = X + \mu\eta$ is equal to the cutoff $k$ believes that the strategic default rate is given by $H(\eta)$ regardless of $\nu$, i.e., by the mass of agents with signal noise lower than hers. But since she does not observe $\eta$, 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 $H(\eta)$), and it is her signal ranking what determines the default rate when $x = k$. This feature of beliefs is known as the Laplacian property (Morris and Shin, 2003) and implies that, while uncertainty about $X$ disappear as noise vanishes, the strategic uncertainty about the default rate does not change. In this scenario, it is easy to see why there cannot be two symmetric equilibria $k, k'$ with $k' > k$: an agent receiving $x = k'$ knows that $X$ is higher than under $x = k$ but holds the same beliefs about $\psi$, leading to $E(P| x = k') > E(P| x = k)$, i.e., to her indifference condition being violated in one of the equilibria.

The argument is more involved when thresholds are asymmetric but a similar logic applies: as we move thresholds up agents beliefs’ about $X$ go up much faster than beliefs about $\psi$, making $E(P| x = k(w))$ to go up.\textsuperscript{24}

Unfortunately, whereas heterogeneity does not affect uniqueness, it complicates the charac-

\textsuperscript{24}The formal proof uses the fact that games with strategic complementarities feature a smallest and largest Nash equilibrium, both in threshold strategies (Milgrom and Roberts, 1990), and the fact that the signal distribution is translation invariant. The latter means that, if an agent receives $x' = x + \Delta$ the distribution of others’ signals conditional on $x'$ is identical to the distribution conditional on $x$ shifted by $\Delta$. Thus, as we shift up cutoffs $k(\cdot)$ by $\Delta$ the distribution of $\psi$ conditional on $k(w) + \Delta$ is identical to its distribution conditional on $k(w)$ before the shift in cutoffs. However, the expected capacity has gone up implying an increase in $E(P| x = k(w))$. Hence, as we move from the smallest to the largest equilibrium, expected monitoring probabilities go up, implying that there must be a unique profile of cutoffs at which $E(P'| x = k(w)) = \theta_{w}(w)$ for all agents with $\theta_{w}(w) \in (0, 1)$. 31
terization of equilibrium as \( \nu \to 0 \). This is because in the presence of heterogeneity the Laplacian property no longer holds and solving for (17) looks hopeless. The reason is that agents with different types use different thresholds and hence the distribution of the strategic default rate conditional on \( x = k(w) \) is no longer uniform. This is not necessarily a problem when solving for \( E(P|x = k(w)) \) when \( k(w) \) is not within \( \nu \) 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 \( \nu \) of each other i.e., those whose thresholds converge to the same limit threshold as \( \nu \downarrow 0 \): we do not know when such a clustering may occur without solving for (17), and to solve these conditions we must know the probability distribution underlying the expectation operator, which can be quite complex. Nonetheless, Sakovics and Steiner (2012) show that the Laplacian property still applies to the average of individual beliefs, as we explain below. We use this remarkable result to circumvent the need to pin down individual beliefs and fully 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. Specifically, we use the joint average condition

\[
\int_{w \in W'} E(P|x = k(w))f(w)dw = \int_{w \in W'} \theta_{\bar{w}}(w)f(w)dw,
\]

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 \( \psi \) by the average belief in the cluster. Since \( X = k(w) \) in the limit when \( x = k(w) \), we can analytically characterize the average expectation in (18) and pin
Sakovics and Steiner (2012) establish that if we average the beliefs of agents 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 $\psi$ 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. While 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 player 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 6 (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.

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

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

$$\frac{1}{\int_{W'} f(w)dw} \int_{W'} \mathbb{P}_w (\psi(W', X) \leq z \mid x = k(w)) f(w)dw = z,$$ (19)
where \( P_w (x = k(w)) \) is the probability assessment of \( \psi(W', X) \) by an agent receiving \( x = k(w) \).

The belief constraint is instrumental to characterize equilibrium thresholds in our model as \( \nu \) goes to zero. As highlighted above, it allows us to express the average indifference condition (18) of types in cluster \( W' \) in a closed form and identify any clusters that may arise in equilibrium. As Proposition 5 states below, the resulting equilibrium threshold is closely related to the shape of \( \theta_w(w)F(w) \). This function represents the level of enforcement capacity \( X \) that makes agents of type \( w \) indifferent between defaulting or not when \( X \) is common knowledge and all agents with returns lower than theirs default (\( \psi = F(w) \)).\(^{25}\) In this regard, Assumption 1 leads to the existence of a unique cluster by ensuring that \( \theta_w(w)F(w) \) is single-peaked, although our characterization technique applies to settings with multiple thresholds (see Theorem 2 in the Appendix).

**Lemma 7.** If Assumption 1 holds then \( \theta_w(w)F(w) \) is increasing at 0 and single-peaked.

Let \( w_{\text{max}} = \arg\max_w \theta_w(w)F(w) \).

**Proposition 5.** In the limit, as \( \nu \to 0 \), there exists \( w^* \geq \bar{w} \) such that

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

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

\[
\theta_w(w^*)F(w^*) (1 - \log \theta_w(w^*)) - F(\bar{w}) = \int_{\bar{w}}^{w^*} \theta_w(w)f(w)dw.
\]

\(^{25}\)Under common knowledge of \( X \), \( \psi \) is common knowledge in equilibrium. Hence, when \( \psi = F(w) \) indifference implies \( P^e = X/F(w) = \theta_w(w) \).
Furthermore, \( \theta_{\bar{w}}(w^*)F(w^*) \) is increasing in \( \bar{w} \).

Figure 3: Equilibrium Strategy in the Global Game.

Figure 3 illustrates the equilibrium strategy implied by Proposition 5. 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, this is not the case on the interval \( w \in [\bar{w}, w^*] \), where agents from the interval cluster on 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 a cluster of agents with types between \( w \) and \( w^* \) simultaneously decide to strategically default.

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-$\theta_w$ 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 $\theta_w$, inducing them to default too. 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 weak 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 $\theta_w(w)F(w)$ being single-peaked, which is 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 $k(w) = \theta_w(w)F(w)$, 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 $\theta_w(w)$. Accordingly, the only way to satisfy indifference conditions is to have a cluster containing agent types on both sides of the peak.\(^{26}\)

A remaining question is that, if all agents in the cluster use the same threshold, how is it possible to simultaneously satisfy their heterogeneous indifference conditions? The answer is that, away from the limit ($\nu > 0$), thresholds of types in the cluster are heterogeneous but are within $\nu$ 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 5 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

\(^{26}\)Multiple peaks due to a multimodal distribution of returns or to non-monotonic default propensities can lead to several clusters. Theorem 2 shows how to characterize them.
propensities of all agents; and an externality effect by increasing the pool of strategic agents in the population. (Here recall that the range of strategic agents is given by \((\bar{w}, \bar{w}/(1 - \gamma \mu))\) by Lemma 2.)

**B. Credit Provision**

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

**Lemma 8.** 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’, i.e., when it is optimal to set credit levels so that \(X\) is enough to prevent the deadweight losses associated to types in the cluster defaulting. This is the case if \(X\) is not too low so that preventing 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.\(^{27}\)

**Proposition 6.** If \(X \geq \theta_{\bar{w}}(w^*)F(w^*)\) at the optimal contract then \(b\) increases with \(X\).

Figure 4 illustrates this result graphically. At a high \(X\), \(b\) is high because there is enough capacity to sustain a high repayment cutoff 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,

\(^{27}\)See footnote 16.
had the contract remained unchanged, a segment of agents between $\bar{w}$ and $\hat{w}$ would default, as indicated in the figure. Since 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.

![Equilibrium Strategy in the Global Game](image)

**Figure 4: Equilibrium Strategy in the Global Game.**

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, i.e., an increase in $\alpha$, results in less capacity buildup and, as a consequence, in lower credit provision. We turn next to consider two concrete numerical examples inspired by the literature that illustrate these effects.
4. Applications

The first application concerns 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. That is, the solution involves the endogenous choice of enforcement capacity $X$ and credit contracts.

A. 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 non-negativity 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 $\gamma$ is arbitrary and set equal to .25. A higher value of $\gamma$ 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.

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.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>1</td>
<td>BGG</td>
</tr>
<tr>
<td>$Ew$</td>
<td>1.02</td>
<td>BGG</td>
</tr>
<tr>
<td>$F$</td>
<td>Lognormal ($\sigma = 3/8$)</td>
<td>BGG, Christiano et al. (2014)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0.88</td>
<td>BGG</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>$c(X)$</td>
<td>0.088X</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Debt-to-equity ($\frac{y}{Ew}$)</td>
<td>0.8</td>
<td>0.5 - 1</td>
</tr>
<tr>
<td>Default rate ($\psi$)</td>
<td>2.3%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Return on Equity (ROE)</td>
<td>3.4%</td>
<td></td>
</tr>
<tr>
<td>Capacity Costs/ROE</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Cluster Size ($Pr(w \in [\bar{w}, w^*])$)</td>
<td>1.7%</td>
<td></td>
</tr>
<tr>
<td>% strategic agents ($\theta_w(w) \in (0, 1)$)</td>
<td>6.9%</td>
<td></td>
</tr>
</tbody>
</table>

---

*BGG refers to Bernanke et al. (1999).*

*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.*

*Christiano et al. (2014) set the debt-to-equity ratio of 0.52, while BGG have it equal to 1.*

*2000-07 Average delinquency rate on business loans. Source: Board of the Federal Reserve.*

Table 1: Parameters and Aggregate Statistics

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 to 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. during the time
period 2000-2007. 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 \( s = 0.018 \) in order 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% prior to the crisis to 4.1% in 2009.\(^{28}\) Although we look at credit provision for different shock probabilities, our calibration assumes that the shock hits with low probability (1%).

Figure 5 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 of \( \gamma = 0 \)). Interestingly, while 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 crucial role in propagating large macroeconomic shocks that increase the rate of non-performing 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.\(^{28}\)

\(^{28}\)Source: Board of Governors of the Federal Reserve System.
Alternative timing. Under alternative timing of events we assume that the principal signs a non-contingent 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 3 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 $s > 0$. What this implies is 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 then the economy will suffer from
inefficiently low credit provision in both normal and distressed times.

B. Propagation of Political Economy Frictions

Our second application concerns an economy-wide 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 $\gamma$ and $\alpha$. 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 my respond in a highly nonlinear way to political economy distortions captured
by preference parameter $\alpha$. To do so we set parameter values similar to the previous setup except
for $R$, and solve the model for different values of $\alpha$ and $\gamma$. 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 6. The horizontal axis shows distortion level $\alpha$, while
the vertical axis reports the supply of credit. Clearly, for low values of $\alpha$ the principal sets $g = 0$ and accumulates as much capacity as possible. Not surprisingly, as $\alpha$ increases the planner diverts more resources for consumption, building less capacity and thus causing a reduction in the credit supply. The interesting result is how the relationship between $\alpha$ and $b$ changes as enforcement externalities become stronger, i.e., as $\gamma$ goes up. Whereas at $\gamma = 0.25$ credit drops in a parsimonious fashion as $\alpha$ goes up, at high $\gamma$ two effects kick in. First, the resource constraint binds at low $\alpha$ making credit unresponsive to small changes in $\alpha$. This is because at higher $\gamma$ the pool of strategic agents is larger (Lemma 2) and thus higher levels of capacity are necessary to sustain a given credit supply. Second, when $\alpha$ is high enough to make principal willing to start diverting resources away from capacity buildup ($g \geq 0$ is no longer binding), a small increase in distortion level $\alpha$ 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 $\gamma$, 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 investment. From a policy perspective, our model suggests that development aid targeting a country’s enforcement infrastructure may be critical in propping up credit markets.

Beyond the context of this particular example, our theory provides an alternative microfoundation of state capacity that highlights the endogeneity of enforcement to the share of non-performing assets in the economy. Under 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 showed that our model can explain two key features characterizing 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 even occur in economies in which there is significant amount of precautionary 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 can magnify the effect political economy distortions on development.
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 $\psi$ and $P$ are also common knowledge. Hence, by Lemma 1, an agent of type $w$ defaults if $P < \theta_w(w)$ 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 $P = \theta_w(\hat{w})$ and the equilibrium monitoring probability is given by

$$P^e = P = \min \left\{ \frac{X}{F(\hat{w})}, 1 \right\}. \quad (21)$$

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

The efficient equilibrium requires $X \geq F(\hat{w})$. Intuitively, this must be the case because agent types with $w$ above but close to $\hat{w}$ 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 $w > \hat{w}$ would have an incentive to default.

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

$$X = \theta_w(\hat{w})F(\hat{w}), \quad (22)$$

This equation admits up to two solutions. This is because, under Assumption 1, $\theta_w(w)F(w)$ is single-peaked (see Lemma 7 below). Figure 7 illustrates the existence of the two inefficient equilibria exhibiting default rates $F(w_2)$ and $F(w_3)$, respectively. Proposition 7 formalizes this result.

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

**Proposition 7.** Under common knowledge of $X$, if $\bar{w} < w_{\text{max}}$, equilibrium is unique iff $X < \underline{X}$ or $X > \bar{X}$. Otherwise, there are three equilibria in the case of $X \in (\underline{X}, \bar{X})$ and two equilibria in the case of $X = \underline{X}$ and $X = \bar{X}$. If $\hat{w} \geq w_{\text{max}}$ equilibrium is always unique.

Appendix II: Proofs

In order 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. We then provide the proofs of the results in the paper by deriving the implications of these results.
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 $\theta_w$ 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 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 $\nu$ sufficiently small.

Note that agents with $w < \bar{w}$ ($\theta(w) > 1$) and those with $\theta(w) < 0$ behave in a non-strategic 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 $0 < \nu < \bar{\nu} := \min\{\theta(w)F(\bar{w}), 1 - \theta(w)\}$. \(^{30}\)

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

**Theorem 1.** The game has an essentially unique equilibrium.\(^{31}\) Equilibrium strategies are characterized by cutoffs $k(w)$ on signal $x$, such that all agents of type $w \in W^*$ 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, i.e., 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, e.g., Milgrom and Roberts (1990) and Vives (1990). Consider the game in which we fix the profile $x$ 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, $a(x)$ and $\bar{a}(x)$ such that any equilibrium profile $a(x)$ satisfies $a(x) \leq a(x) \leq \bar{a}(x)$. Moreover, if we fix the action profile of all agents, the difference in expected payoff from choosing $a = 1$ versus $a = 0$ for any given agent is increasing in $x$, 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. $x$, and Theorem 6 in Milgrom and Roberts (1990) applies: $a(x)$ and $\bar{a}(x)$ are non-decreasing functions of $x$. But since 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 to establish the following translation result: when all cutoffs are shifted by the same amount $\Delta$ expected monitoring probabilities go up. Equipped with such result 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 (23) are satisfied.

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

**Lemma 9.** If $k$ is a profile of equilibrium cutoffs then $k(w) \in [(\theta(w) - \nu/2)F(\bar{w}), \theta(w) + \nu/2]$ for all $w \in W^*$.

---

\(^{30}\)This upper bound on $\nu$ 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.

\(^{31}\)In the sense that equilibrium strategies may differ in zero probability events.
Hence, we can obtain the following inequality by a well-defined change of variables:

$$\mathbb{E}_\theta[P|k; k(w)] = \theta(w) \quad \forall w \in \mathcal{W}^*. \quad (23)$$

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 (21) we have that, if $k(w) \in [\nu/2, 1 - \nu/2]$,

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

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

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

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

**Lemma 10.** If $k$ is a profile of equilibrium cutoffs then $\mathbb{E}[P|k; k(w)] < \mathbb{E}[P|k + \Delta; k(w) + \Delta]$ 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 $\nu \leq \theta(w)F(\bar{w}) \leq \theta(w)F(\bar{w})$ and, by Lemma 9, $k(w) \geq (\theta(w) - \nu/2)F(\bar{w})$ we have that $k(w) \geq \nu/2$. Likewise, $k(w) + \Delta \leq \bar{k}(w) \leq 1 - \nu/2$ by Lemma 9 and the fact that $\nu \leq 1 - \theta(w_i)$. Hence, we can obtain the following inequality by a well-defined change of variables:

$$\mathbb{E}[P(X)|k; k(w)] =$$

$$\int_{-1/2}^{1/2} \min \left\{ \frac{X}{F(\bar{w}) + \sum w' H \left( \frac{k(w') - X}{\nu} \right) f(w')} \cdot 1 \right\} h \left( \frac{k(w) - X}{\nu} \right) dX$$

$$< \int_{-1/2}^{1/2} \min \left\{ \frac{X + \Delta}{F(\bar{w}) + \sum w' H \left( \frac{k(w') - X}{\nu} \right) f(w')} \cdot 1 \right\} h \left( \frac{k(w) - X}{\nu} \right) dX$$

$$= \int_{-1/2}^{1/2} \min \left\{ \frac{X' + \Delta - X'}{F(\bar{w}) + \sum w' H \left( \frac{k(w') + \Delta - X'}{\nu} \right) f(w')} \cdot 1 \right\} h \left( \frac{k(w) + \Delta - X'}{\nu} \right) dX'$$

$$= \mathbb{E}[P|k + \Delta; k(w) + \Delta].$$

The inequality is strict since $k$ being an equilibrium profile means that $\mathbb{E}[P|k; k(w)] = \theta(w) < 1$ for all $w \in \mathcal{W}^*$. Accordingly, monitoring probabilities, conditional on $x = k(w)$, are less than
one 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 \( \Delta \).

Equipped with Lemma 10 we know argue that \( k = \bar{k} \). Assume, by way of contradiction, that \( \hat{k}(w) < \bar{k}(w) \) for some \( w \in \mathcal{W}^* \). Denote \( \hat{w} = \arg\max_{w \in \mathcal{W}^*} (\bar{k}(w) - k(w)) \) and \( \hat{\Delta} = \bar{k}(\hat{w}) - \hat{k}(\hat{w}) \). By Lemma 10, we have that

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

where the last inequality comes from the fact that default rates at \( \bar{k} \) are lower than at \( k + \hat{\Delta} \geq \bar{k} \), 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 (23).

In order to solve this system of equations we need to pin down agent beliefs when they receive their threshold signals. In order 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

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

**Lemma 11 (Belief Constraint).** For any subset \( W' \subseteq \mathcal{W}^* \) and any \( z \in [0, 1] \),

\[
\frac{1}{\sum_{W'} f(w)} \sum_{W'} \mathbb{P}(\psi(X, W') \leq z | x = k(w)) f(w) = z.
\]

**Proof.** The result follows directly from the proof of Lemma 1 in Sakovics and Steiner (2012). To see why note first that Lemma 9 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 \( w \in W' \) and focus on the aggregate action of agents with types in \( W' \), rather than the aggregate action in the population.\(^{32}\)

The above result is instrumental to characterize equilibrium thresholds as \( \nu \) goes to zero. In particular, it allows to derive closed-form solutions for the above indifference conditions from which we can obtain \( k \). 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 \( F^{-}(w) = \sum_{w' < w} f(w') \).

\(^{32}\)When thresholds do not fall within \( \nu \) of each other, the distribution \( \hat{F} \) of virtual errors \( \hat{\eta} \) need not be strictly increasing and thus its inverse may not be well-defined. Defining \( \hat{F}^{-1}(u) = \inf\{ \hat{\eta} : \hat{F}(\hat{\eta}) \geq u \} \) takes care of this issue and ensures that the proof of Lemma 2 in Sakovics and Steiner (2012) applies to the general case.
Theorem 2. In the limit, as \( \nu \to 0 \), the equilibrium cutoff strategies are given by a unique monotone partition \( \Phi = \{W_1, \ldots, W_I\} \) and a unique vector \((k_1, \ldots, k_I)\) satisfying the following conditions:

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

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

(iii) \( \theta(w_i) F^-(w_i) \leq k_i \leq \theta(\bar{w}_i) F(\bar{w}_i) \) for all \( i = 1, \ldots, I \).

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

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 to \( \nu > 0 \), with \( k^\nu \) denoting the equilibrium cutoff profile. The first step of the proof is to show that \( k^\nu \) uniformly converge as \( \nu \to 0 \) and identify the set of indifference conditions that pin down the limit equilibrium. Let

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

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

Lemma 12. There exists a unique partition \( \{W_1, \ldots, W_I\} \) and a set of thresholds \( k_1 > k_2 > \cdots > k_I \) such that, as \( \nu \to 0 \), for all \( w \in W_i \), \( i = 1, \ldots, I \), \( k^\nu(w) \) uniformly converges to \( k_i \). Moreover, thresholds \( k = (k_1, \ldots, k_I) \) solve the system of limit indifference conditions

\[
\int_0^1 \min \left\{ \frac{k_i}{F(\bar{v}) + \sum_{W_i \cup_j \leq W_j} f(w') + z \sum_{W_i} f(w')} + 1 \right\} dA_w(z|k, W_i) = \theta(w), \quad \forall w \in W_i, \forall i,
\]

where \( A_w(z|k, W_i) \) represents the strategic beliefs of type-\( w \) agents in the limit and satisfies the belief constraint (25).

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 \( w > \bar{w} \) such that \( w \in W_i \) and \( \bar{w} \in W_m \) with \( m > i \). First note that the LHS in (26) is bounded below by \( \frac{k_i}{F(\bar{v}) + \sum_{W_i \cup_j \leq W_j} f(w')} \) and bounded above by \( \min \left\{ k_i / F(\bar{v}) + \sum_{W_i \cup_j \leq W_j} f(w'), 1 \right\} \).

Given this, since \( \theta(\bar{w}) < 1 \) the monitoring probability when all agents with types in \( W_m \) default is strictly less than one, i.e., \( \frac{k_m}{F(\bar{v}) + \sum_{W_m \leq W_j} f(w')} < 1 \). Otherwise (26) would be violated. In
addition, $m > i$ implies that $k_m < k_i$ by the above lemma and that $\sum_{j<i} f(w') < \sum_{j\leq m} f(w')$. Combining all this we arrive to the following contradiction

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

The monotonicity of the type partition implies that $F(\bar{w}) + \sum_{j<i} f(w') = F(\bar{w}_i)$ and that $F(\bar{w}) + \sum_{j>i} f(w') = F^- (\bar{w}_j)$. Given this, it is straightforward to check that the above bounds on the LHS of (26) lead to condition (iii) in the theorem.

Finally, in order to obtain condition (iv) from (26) we make use of the belief constraint in the limit, which can be written as

$$\frac{1}{\sum_{W_i} f(w)} \sum_{W_i} A_w(z|k, W_i) f(w) = z. \quad (27)$$

Multiplying both sides of (26) by $\frac{f(w)}{\sum_{W_i} f(w)}$ and summing over $w \in W_i$ we get

$$\int_0^1 \min \left\{ \frac{k_i}{F^- (w_i) + z \sum_{W_i} f(w')}, 1 \right\} d \left( \frac{1}{\sum_{W_i} f(w)} \sum_{W_i} A_w(z|k, W_i) f(w) \right) = \sum_{W_i} \theta(w) f(w) \sum_{W_i} f(w). \quad (28)$$

Finally, using the belief constraint (27) to substitute for the last term in the LHS we obtain

$$\int_0^1 \min \left\{ \frac{k_i}{F^- (w_i) + z \sum_{W_i} f(w')}, 1 \right\} dz = \frac{\sum_{W_i} \theta(w) f(w)}{\sum_{W_i} f(w)}. \quad (28)$$

Note that $F^- (w_i) + z \sum_{W_i} f(w') \sim U[F^- (w_i), F(\bar{w}_i)]$ with density $\frac{1}{f(w)}$ since $z \sim U[0, 1]$. Hence, we can rewrite (28) as

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

yielding condition (iv).

\[ \square \]

**Proof of Lemma 12.** To prove convergence we first partition the set of types into subsets $W_i$ of types for sufficiently small $\nu$ as follows: (i) if we order the signal thresholds of all types, any adjacent thresholds that are within $\nu$ 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 \( \nu \). Also, let \( Q_w^\nu (x|k^\nu, z) := \mathbb{P} \left( X \leq x \mid x = k^\nu (w), \psi(X, W_i) = z \right) \) 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 \( Q_w^\nu (|k^\nu, z) \) 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:

\[
\mathbb{E}(P|k^\nu; k^\nu (w)) = \\
\int_0^{1} \int_{k^\nu (w) - \nu/2}^{k^\nu (w) + \nu/2} \min \left\{ \frac{F(\bar{w}) + \sum_{w_j \in W_j} \frac{X}{f(w')} + z \sum_{W_i} f(w')}{1} \right\} \ dQ_w^\nu (x|k^\nu, z) dA_w(z|k^\nu, W_i). \tag{29}
\]

In addition, notice that we can always express \( \mathbb{E}(P|k^\nu; k^\nu (w)) \) 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 \( \Delta_{W_i} \) fixed, \( A_w(z|k^\nu, W_i) \) does not change with \( \nu \). 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 also fix \( \Delta_{W_i} \), for all \( i = 1, \ldots, I \) and all \( \nu \) 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 \( \mathbb{E}(P|k^\nu; k^\nu (w)) = \theta (w) \) is approximated by the limit condition in the lemma for \( \nu \) sufficiently small.

Note that the inner integral in (29) is bounded below by \( \min \left\{ \frac{k^\nu (w) - \nu/2}{F(\bar{w}) + \sum_{w_j \in W_j} f(w') + z \sum_{W_i} f(w')} \right\} \) and above by \( \left\{ \frac{k^\nu (w) + \nu/2}{F(\bar{w}) + \sum_{w_j \in W_j} f(w') + z \sum_{W_i} f(w')} \right\} \). Hence,

\[
\int_0^{1} \min \left\{ \frac{k_i - \nu/2}{F(\bar{w}) + \sum_{w_j \in W_j} f(w') + z \sum_{W_i} f(w')} \right\} dA_w(z|k^\nu, W_i) \leq \mathbb{E}(P|k^\nu; k^\nu (w)) \leq \int_0^{1} \min \left\{ \frac{k_i + \nu/2}{F(\bar{w}) + \sum_{w_j \in W_j} f(w') + z \sum_{W_i} f(w')} \right\} dA_w(z|k^\nu, W_i). \tag{30}
\]

The first term in these integrals is Lipschitz continuous. In addition, the next lemma shows

---

\footnote{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 (21), we find that \( \psi(X, W_i) \) only depends on \( \Delta_{W_i} \) and \( k^\nu (w) \). But this means that \( A_w(z|k^\nu, W_i) \) only depends on \( \Delta_{W_i} \) and \( k^\nu (w) \) since \( h \) is independent of \( \nu \).}

54
that $dA_w(z|k^\nu, k^\nu(w))$ is bounded for all $\nu$.

**Lemma 13.** $0 \leq \frac{\partial A_w(z|k^\nu, k^\nu(w))}{\partial z} \leq \frac{\sum w_i f(w')}{f(w)}$ for all $w \in W_i$ and all $z$ in the support of $A_w(\cdot|k^\nu, k^\nu(w))$.

See proof below.

Hence, the LHS and the RHS of (30) uniformly converge to each other as $\nu \to 0$, leading to limit indifference conditions (26). 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 $E(P|k^\nu; k^\nu(w)) = \theta(w)$ lies in a compact set.\(^{34}\) Accordingly, we can find $\hat{\nu}$ so that indifference conditions are within $\varepsilon$ of the limit condition for all $\nu < \hat{\nu}$, leading to their solutions being in a neighborhood of the solution $k$ of limit indifference conditions (26).

**Proof of Lemma 13.** 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 $\psi^{-1}$ 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

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

Differentiating w.r.t. $z$ yields

$$\frac{\partial A_w(z|k^\nu, W_i)}{\partial z} = \frac{1}{\nu} h\left(\frac{k^\nu(w) - \psi^{-1}(z, W_i)}{\nu}\right) \left(-\frac{\partial \psi^{-1}(z, W_i)}{\partial z}\right)$$

$$= \frac{1}{\sum w_i f(w')} \sum w_i H\left(\frac{k^\nu(w') - \psi^{-1}(z, W_i)}{\nu}\right) f(w').$$

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

**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.

\(^{34}\)If $\{W_i\}_{i=1}^I$ is not kept fixed then when $\nu$ is very small $E(P|k^\nu; k^\nu(w))$ would be discontinuous at some $\nu$, implying a violation of the indifference condition for some $w \in W^*$.  

55
Proof of Lemma 1. The propensity to default $\theta$ 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, $\theta$ 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 4. As stated in text, uniqueness should be interpreted as the existence of a unique equilibrium in nearby discrete-return economies (Theorem 1).

Proof of Lemma 6. The statement of the lemma is a continuous-returns version of Lemma 11.

Proof of Lemma 7. 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 $\frac{F(w)}{wf(w)}$ is increasing the expression is strictly decreasing. Now, since the second term is zero at $w = 0$ and $\frac{F(w)}{wf(w)}$ is increasing and $\lim_{w \downarrow 0} \frac{F(w)}{wf(w)} < 1$ 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 5. In order 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 $\{(\bar{w}_j, \bar{w}_{j+1})\}_{j=1}^I$ 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 \( w \in (\bar{w}_j, \tilde{w}_j) \);

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

\[
\int_{F(\tilde{w}_j)}^{F(\bar{w}_j)} \min \left\{ \frac{k_i}{z}, 1 \right\} \,dz = \int_{\bar{w}_j}^{\tilde{w}_j} \theta(w)f(w)\,dw. \tag{31}
\]

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)-(b) and conditions (iii)-(iv). Since \( \theta(w)F(w) \) is continuous parts (a) and (b) imply that \( k(\tilde{w}) = \theta(w)F(w) \) at the boundaries of an interval in which \( k \) is constant, except possibly when \( w_i = \tilde{w} \), in which case condition (iii) requires that \( k_j \geq \theta(\bar{w}_j)F(\bar{w}_j) \). Expression (31) is the continuous counterpart of (iv).

We now argue that the single-peakedness of \( \theta(w)F(w) \) (Lemma 7 in Appendix I) leads to a partition consisting of two intervals, the first one where \( k \) is constant and the second one in which it is strictly decreasing.

First notice that \( k \) being decreasing implies that there must be at least one pooling threshold since \( \theta(w)F(w) \) is initially increasing. To show why there is only one we use the fact that condition (c) requires that \( k_1 = \theta(\tilde{w}_1)F(\tilde{w}_1) \geq \theta(\bar{w}_1)F(\bar{w}_1) \). 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 \( \tilde{w}_1 \) and decreasing at \( \bar{w}_1 \). Otherwise, either \( \theta(w)F(w) \) is decreasing at \( \tilde{w}_1 \) or \( \theta(w)F(w) \) is increasing in \( (\bar{w}_1, \tilde{w}_1) \). The former case implies that \( \theta(\tilde{w}_1)F(\tilde{w}_1) < \theta(\bar{w}_1)F(\bar{w}_1) \), violating (c). The latter case implies that \( k_1 = \max_{w \in [\bar{w}_1, \tilde{w}_1]} \theta(w)F(w) \), which implies that the LHS of (31) is greater than the RHS.

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

We finish the characterization of equilibrium thresholds by showing that \( \tilde{w}_1 = w^* \), where \( w^* \) is the unique solution to (20) in \( (\tilde{w}, \infty) \) when \( \tilde{w} < w_{max} \).

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

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

where the last equality comes from the change in variable \( w = F^{-1}(z) \) (\( dz = f(w)\,dw \)).
\[ k_1 = \theta(w^*)F(w^*) \text{ and } w_1 = \bar{w} \text{ we can express the LHS of (31) as} \]
\[
\int_{k_i}^{F(w^*)} \frac{k_i}{z} dz + \int_{F(\bar{w})}^{k_i} dz = k_i \log \left( \frac{F(w^*)}{k_i} \right) + k_i - F(\bar{w}) = \theta(w^*)F(w^*)(1 - \log \theta(w^*)) - F(\bar{w}).
\]

Equating the RHS of the last expression to the RHS of (31) yields (20). To show that it has a unique solution in \((\bar{w}, \infty)\) we express it as
\[
\theta(w^*)F(w^*)(1 - \log \theta(w^*)) - \int_{\bar{w}}^{w^*} \theta(w)f(w)dw = F(\bar{w}), \tag{32}
\]
and differentiate the LHS w.r.t. \(w^*\), which yields \((- \log \theta(w^*)) (\theta'(w^*)F(w^*) + \theta(w^*)f(w^*))\). 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 \(\theta\) 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 (32):
\[
\frac{\partial \text{LHS}}{\partial w^*} \frac{dw^*}{d\bar{w}} = \int_{\bar{w}}^{w^*} \frac{\partial \theta(w)}{\partial \bar{w}} f(w)dw.
\]

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

**Proof of Lemma 8.** To prove that the principal always sets \(b < \infty\), first note that Proposition 5 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:
\[
\frac{b}{y + b} \leq \int_{\hat{w}}^{\infty} \hat{w}dF(w) + \mu (P + (1 - \gamma)(1 - P)) \int_{0}^{\hat{w}} wdF(w). \tag{33}
\]

The LHS converges to one as \(b \to \infty\). Hence, we need to show that the RHS is strictly less than one for all \(\hat{w}\). Since \(\hat{w} \geq w\) and the RHS is increasing in \(P\) for fixed \(\hat{w}\), the RHS is bounded above by
\[
\int_{\hat{w}}^{\infty} \hat{w}dF(w) + \mu \int_{0}^{\hat{w}} wdF(w), \tag{34}
\]
which is strictly less than one for all \(\hat{w}\) by Assumption 2. \(\square\)
Proof of Proposition 6. 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 5. 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 \( X \geq \theta_{\bar{w}}(w^*)F(w^*) \).\(^{36}\)

To show that \( b \) goes up with \( \bar{w} \), we first note that the budget constraint (33) 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 \( b \). Since the LHS of budget constraint (33) is strictly increasing in \( b \) that means that, 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 to \( X \). If that is the case then a higher capacity relaxes the constraint on \( \bar{w} \), in turn relaxing the constraint (33) 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 to \( X \). 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.\(^{37}\) Accordingly, \( b \) must be increasing in \( \bar{w} \) at the optimal contract associated to \( X \).

Finally, if \( X' < \theta_{\bar{w}'}(w^{**})F(w^{**}) \) it must be that \( \bar{w}' > \bar{w} \) since \( \theta_{\bar{w}'}(w^{**})F(w^{**}) > \theta_{\bar{w}}(w^*)F(w^*) \). 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. \( \square \)

\(^{36}\) 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.

\(^{37}\) Formally, agents’ payoffs are given by \( \int_{\bar{w}}^{\infty} (y + b)(w - \bar{w})dF + P\gamma\mu \int_{\bar{w}}^{\infty} (y + b)\bar{w}dF \). Since the budget constraint holds with equality, we can express the last term as

\[
P\gamma\mu \int_{0}^{\bar{w}} (y + b)\bar{w}dF = b - \int_{\bar{w}}^{\infty} (y + b)\bar{w}dF - (1 - \gamma)\mu \int_{0}^{\bar{w}} (y + b)\bar{w}dF.
\]

Hence, agents’ payoffs can be written as

\[
\int_{\bar{w}}^{\infty} (y + b)\bar{w}dF + b - (1 - \gamma)\mu \int_{0}^{\bar{w}} (y + b)\bar{w}dF,
\]

which strictly go up if we increase \( b \) and lower \( \bar{w} \).
Proof of Proposition 7 in Appendix I. First consider the case of \( X < \bar{X} \). If \( \bar{w} < w_{\text{max}} \) it must be that (20) has only solutions, \( w_1 < \bar{w} \) and \( w_2 > w_{\text{max}} \). 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 \( \bar{w} \geq w_{\text{max}} \).

When \( X = \bar{X} \) 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 \( w_{\text{max}} \) is a solution of (20), representing a second equilibrium.

Finally, if \( X \in (\bar{X}, \bar{X}) \) there are three equilibria given by the two solutions in \([\bar{w}, \infty)\) to (20) and another equilibrium with \( \psi = F(\bar{w}) \) since \( F(\bar{w}) < X \) and thus the principal can credibly sustain \( P = 1 \) in equilibrium.

\[ \square \]

References


Woo, David, “Two Approaches to Resolving Nonperforming Assets During Financial Crises,”