Crossing the Threshold: Why Finance and Growth Diverge

Lei Pan*
Curtin University

Abstract

Why do economies with similar fundamentals display persistently different levels of financial development and growth? This paper builds a three-sector continuous-time general-equilibrium model in which financial development is an endogenous stock that raises allocative efficiency, while intermediation faces fixed operating costs and congestion losses. The interaction generates multiple steady states: a low-finance trap, an unstable threshold, and a high-finance regime. A simple phase diagram clarifies the mechanism and policy levers that eliminate traps.

Keywords: Financial development; Multiple steady states; Growth traps; Entry costs; Allocative efficiency

JEL Classification: E22; E44; O16; O41

*School of Accounting, Economics and Finance, Curtin University, Perth, Australia. Email: lei.pan@curtin.edu.au

1 Introduction

Finance is often described as the economy's plumbing: when it works, funds reach productive users quietly; when it fails, even good ideas struggle to get off the ground. A simple puzzle follows. Many economies start with similar savings motives and similar technologies, yet their financial systems—and therefore their growth paths—can diverge for long periods. One economy gradually builds reliable intermediation, deeper markets, and better screening; another remains stuck with thin intermediation, high per-unit costs, and fragile trust. This paper offers a parsimonious mechanism for such divergence: when intermediation has fixed operating costs and congestion costs, financial development can become self-reinforcing only after the system reaches a critical mass.

The idea that finance shapes development is old, going back to (Goldsmith 1969, McKinnon 1973, Shaw 1973). Modern empirical work documents a strong association between financial depth and long-run growth and its sources (King and Levine 1993, Levine 1997, Rajan and Zingales 1998, Beck *et al*. 2000), while a large institutional literature stresses that contract enforcement and investor protection are central to financial performance (La Porta *et al*. 1998, Acemoglu *et al*. 2005). On the theory side, classic models link finance and growth through costly intermediation and endogenous adoption of financial structures (Greenwood and Jovanovic 1990, Bencivenga and Smith 1991). At the same time, recent evidence suggests nonlinearities: beyond some point, finance can become less helpful or even harmful (Arcand *et al*. 2015, Cecchetti and Kharroubi 2012).

Against this background, we develop a three-sector continuous-time general-equilibrium model with households, producers, and financial intermediaries. Financial development is a stock that raises allocative efficiency, but it is produced endogenously through intermediation activity, entry, and learning. The model delivers multiple steady states: a low-finance trap, a high-finance regime, and an unstable threshold separating them. The analysis yields three contributions. First, it provides a clean state-variable representation of financial deepening that nests both "finance helps growth" and threshold-style dynamics in one setup. Second, it shows how fixed operating costs and congestion in intermediation generate multiplicity in a transparent way, with a simple two-dimensional phase diagram. Third, it highlights policy levers with clear interpretation: lowering operating costs, improving screening effectiveness, or raising the productivity of deepening can eliminate the trap by shifting the deepening locus and making the high-finance steady state globally attractive.

The remainder of the paper proceeds as follows. Section 2 presents the three-sector model and defines equilibrium. Section 3 concludes.

2 Model

This section builds a three-sector continuous-time general-equilibrium model—households, producers, and financial intermediaries—with: i) endogenous financial deepening, ii) intermediary entry with fixed operating costs, iii) screening effort and congestion in intermediation, and iv) an explicit nonlinearity that delivers multiple steady states.

2.1 Set up

Time is continuous, $t \in [0, \infty)$. The economy is closed. A single final good can be consumed or invested. Population is normalized to one and supplies one unit of labor inelastically.[1]

There are three sectors: i) a representative household that chooses consumption and saving dynamically; ii) a competitive final-good sector that uses effective capital to produce output; and iii) a financial intermediation sector that screens and channels resources, and endogenously accumulates a stock of "financial development" that improves allocative efficiency.

The representative household has CRRA utility over consumption:

$$\max_{\{C_t\}_{t\geq 0}} \int_0^\infty e^{-\rho t} \frac{C_t^{1-\gamma} - 1}{1-\gamma} \, dt, \tag{1}$$

where $\rho > 0$ is the subjective discount rate and $\gamma > 0$ is the inverse of the intertemporal elasticity of substitution.

Let K_t denote the economy-wide stock of physical capital owned by the household. Capital depreciates at rate $\delta > 0$. The household receives the rental return R_t on capital and the wage w_t from inelastic labor supply. The household also receives aggregate profits from intermediaries, denoted Π_t, which will be zero under free-entry/competition in the baseline equilibrium.

Household capital accumulation is:

$$\dot{K}_t = Y_t - C_t - C_F(N_t, e_t, F_t) - \delta K_t, \tag{2}$$

where Y_t is aggregate output, and $C_F(\cdot)$ is the real resource cost of running the financial system (labor and goods absorbed by screening, operating costs, and intermediation frictions). The term C_F is the key channel through which finance both helps and (potentially) drags on the economy: it raises efficiency via F_t (defined below), but also uses resources.

[1] Extensions with population growth are straightforward and omitted for clarity.

Let Λ_t denote the costate variable on Equation (2). The household's Hamiltonian is:

$$\mathcal{H}_t = \frac{C_t^{1-\gamma} - 1}{1-\gamma} + \Lambda_t\Big(Y_t - C_t - C_F(N_t, e_t, F_t) - \delta K_t\Big). \tag{3}$$

The first-order condition for consumption is:

$$\frac{\partial \mathcal{H}_t}{\partial C_t} = 0 \quad \Rightarrow \quad C_t^{-\gamma} = \Lambda_t. \tag{4}$$

The costate equation is:

$$\dot{\Lambda}_t = \rho\Lambda_t - \frac{\partial \mathcal{H}_t}{\partial K_t} = (\rho + \delta)\Lambda_t - \Lambda_t \frac{\partial Y_t}{\partial K_t}, \tag{5}$$

where we use that $\partial C_F / \partial K_t = 0$ (below we allow C_F to depend on activity and thus indirectly on K_t through Y_t and C_t).

Combining Equation (4)–(5) yields the Euler equation:

$$\frac{\dot{C}_t}{C_t} = \frac{1}{\gamma}\left(\frac{\partial Y_t}{\partial K_t} - \delta - \rho\right). \tag{6}$$

Equation (6) matches the logic that consumption growth is pinned down by the net marginal product of capital.

The final-good sector is competitive. Output is produced using capital, but the productivity of capital depends on a stock of financial development F_t:

$$Y_t = A\,\mathcal{A}(F_t)\,K_t, \tag{7}$$

where $A > 0$ is baseline productivity and $\mathcal{A}(F)$ is an increasing, bounded efficiency term capturing allocative efficiency, contract enforcement, information, and payment infrastructure.

A parsimonious functional form that is smooth, increasing, and saturating is:

$$\mathcal{A}(F) = \left(\frac{F}{\bar{F} + F}\right)^{\vartheta}, \qquad \bar{F} > 0,\ \vartheta > 0. \tag{8}$$

When $F \ll \bar{F}$, efficiency is low and marginal improvements in F can matter a lot; when F is very large, efficiency approaches one and gains taper off.

Under Equation (7), the marginal product of capital is:

$$\frac{\partial Y_t}{\partial K_t} = A\,\mathcal{A}(F_t). \tag{9}$$

With inelastic labor normalized to one, wages can be interpreted as a residual share if desired; for the present purposes, Equation (9) is the key object because it enters Equation (6).

Substituting Equation (9) into Equation (6) delivers:

$$\frac{\dot{C}_t}{C_t} = \frac{1}{\gamma}\left(A\,\mathcal{A}(F_t) - \delta - \rho\right). \tag{10}$$

Hence, financial development affects intertemporal decisions by shifting the effective return to saving.

The financial sector is modelled as an industry with $N_t \geq 0$ active intermediaries at time t. Intermediaries provide screening/monitoring and payment/verification services. These services i) absorb resources today but ii) raise the stock F_t that improves allocative efficiency in Equation (7).

Let $e_t \geq 0$ denote average screening/monitoring intensity per intermediary. The financial sector absorbs real resources according to:

$$C_F(N_t, e_t, F_t) = \underbrace{f\,N_t}_{\text{fixed operating cost}} + \underbrace{\frac{\kappa}{2}\,e_t^2\,\mathcal{S}_t}_{\text{screening cost}} + \underbrace{\chi\,\Phi\!\left(\frac{\mathcal{S}_t}{N_t + \underline{N}}\right)\mathcal{S}_t}_{\text{congestion / misallocation loss}}, \tag{11}$$

where $f > 0$ is a fixed operating cost per intermediary (branches, compliance, IT overhead); $\kappa > 0$ scales the marginal cost of screening intensity; $\mathcal{S}_t$ is the scale of financial activity to be intermediated. A natural object is gross saving/investment resources, $\mathcal{S}_t \equiv Y_t - C_t$, i.e., resources not consumed are the flow that must be allocated across projects and firms. $\underline{N} > 0$ prevents singularity at $N_t = 0$ and captures that some basic "backbone" (courts, registry) may exist even with few intermediaries; and $\Phi(\cdot)$ captures congestion: intermediating a given scale with few intermediaries raises per-intermediary load and worsens screening/verification quality.

A tractable congestion function is:

$$\Phi(x) = x^{\nu}, \qquad \nu > 0. \tag{12}$$

Then the congestion term in Equation (11) scales like $\mathcal{S}_t^{1+\nu}/(N_t + \underline{N})^{\nu}$.

Financial development F_t evolves with experience and infrastructure investment. It rises with intermediation scale and with the number of active intermediaries (entry and competition bring new technologies and networks), but depreciates as institutions and systems become obsolete:

$$\dot{F}_t = \xi\left(\frac{N_t}{N_t + \bar{N}}\right)^{\omega}\left(\frac{\mathcal{S}_t}{Y_t}\right)^{\psi} F_t\left(1 - \frac{F_t}{F^{\max}}\right) - \delta_F F_t, \tag{13}$$

where $\xi > 0$ is the speed of financial deepening (how quickly learning-by-doing and infrastructure translate into higher F); $\bar{N} > 0$ normalizes how rapidly the "competition/network" channel saturates; $\omega > 0$ governs how strongly entry and network size matter for deepening; $\psi > 0$ governs how strongly the saving rate (intermediation intensity) fuels deepening; $F^{\max} > 0$ is an upper bound, introducing a natural saturation effect; and $\delta_F > 0$ is institutional depreciation (obsolescence, erosion of trust, regulatory decay).

Two features of Equation (13) are important for multiplicity. First, deepening is state-dependent via $F_t(1 - F_t/F^{\max})$, so very low F is hard to escape. Second, deepening depends on the endogenous choice/entry N_t and on the endogenous saving rate $\mathcal{S}_t/Y_t$.

To close the model, N_t is determined endogenously. A simple and empirically plausible closure is free entry with zero expected profits. Rather than modelling the full pricing problem (which is not needed for the core multiplicity mechanism), we represent the gross revenue generated by intermediation as proportional to the scale of intermediated funds and increasing in F_t:

$$\mathcal{R}_t = \mu\, \mathcal{B}(F_t)\, \mathcal{S}_t, \qquad \mu > 0, \tag{14}$$

with $\mathcal{B}(F)$ increasing and bounded, for example

$$\mathcal{B}(F) = \left(\frac{F}{\bar{F} + F}\right)^{\eta}, \qquad \eta > 0. \tag{15}$$

The interpretation is that when F is low, contracts are hard to enforce and informational frictions are large; intermediated flows generate limited effective revenue and fee base. When F is high, the system is more reliable, and intermediation activity generates a larger feasible fee base (more transactions, better recoveries, more scalable services).

Under free entry, aggregate revenue (14) is dissipated into operating and screening costs. A reduced-form zero-profit condition is:

$$\mathcal{R}_t = C_F(N_t, e_t, F_t). \tag{16}$$

Using Equation (11)–(14), entry pins down N_t implicitly as a function of (K_t, C_t, F_t, e_t).

Effort e_t can be taken as a policy/institutional choice (regulatory intensity) or as an industry choice. A convenient closure is that e_t is chosen to minimize per-unit intermediation losses, trading off screening cost against congestion losses. Formally, given (N_t, F_t) and $\mathcal{S}_t$, choose e_t to minimize the variable part of Equation (11):

$$\min_{e_t \geq 0} \; \frac{\kappa}{2} e_t^2 \mathcal{S}_t \; + \; \chi \left(\frac{\mathcal{S}_t}{N_t + \underline{N}}\right)^{\nu} \mathcal{S}_t, \tag{17}$$

which implies $e_t^\star = 0$ in this stripped-down form (because congestion losses do not depend on e_t directly). To make effort meaningful, allow effort to reduce effective congestion:

$$\Phi\left(\frac{\mathcal{S}_t}{N_t + \underline{N}}\right) = \left(\frac{\mathcal{S}_t}{N_t + \underline{N}}\right)^{\nu} \exp(-\zeta e_t), \qquad \zeta > 0. \tag{18}$$

Then the effort FOC yields an interior solution:

$$\kappa e_t = \chi\, \zeta \left(\frac{\mathcal{S}_t}{N_t + \underline{N}}\right)^{\nu} \exp(-\zeta e_t), \tag{19}$$

which has a unique solution because the LHS is increasing in e_t and the RHS is decreasing in e_t. Equation (19) captures that: i) larger scale $\mathcal{S}_t$ raises the marginal benefit of monitoring, ii) more intermediaries (larger N_t) reduces load and thus reduces the need for very high effort, and iii) the parameter ζ governs how effective monitoring is at mitigating misallocation.

In what follows, we take $e_t = e^\star(\mathcal{S}_t, N_t)$ as implicitly defined by Equation (19).

2.2 Competitive equilibrium

A competitive equilibrium is a set of allocations and prices

$$\{C_t, K_t, Y_t, F_t, N_t, e_t, R_t, w_t\}_{t\geq 0}$$

such that i) given prices and laws of motion, the household maximizes Equation (1) subject to Equation (2), implying the Euler equation (10) and transversality; ii) the production sector satisfies Equation (7)–(9); iii) financial development evolves according to Equation (13); iv) intermediary entry satisfies Equation (16), with effort satisfying Equation (19); and v) goods market clearing holds by construction in Equation (2) with $\mathcal{S}_t = Y_t - C_t$.

2.3 Reduced system and multiple steady states

Because Y_t is linear in K_t in Equation (7), it is convenient to work with the consumption-capital ratio

$$x_t \equiv \frac{C_t}{K_t}. \tag{20}$$

Using Equation (7), output per unit of capital is:

$$\frac{Y_t}{K_t} = A\, \mathcal{A}(F_t). \tag{21}$$

The saving flow per unit of capital is:

$$\frac{\mathcal{S}_t}{K_t} = \frac{Y_t - C_t}{K_t} = A\mathcal{A}(F_t) - x_t. \tag{22}$$

Dynamics of x_t. Differentiate $x_t = C_t/K_t$:

$$\frac{\dot{x}_t}{x_t} = \frac{\dot{C}_t}{C_t} - \frac{\dot{K}_t}{K_t}. \tag{23}$$

From Equation (10),

$$\frac{\dot{C}_t}{C_t} = \frac{1}{\gamma}\left(A\mathcal{A}(F_t) - \delta - \rho\right). \tag{24}$$

From Equations (2) and (21),

$$\frac{\dot{K}_t}{K_t} = A\mathcal{A}(F_t) - x_t - \frac{C_F(N_t, e_t, F_t)}{K_t} - \delta. \tag{25}$$

Hence,

$$\dot{x}_t = x_t\left[\frac{1}{\gamma}\left(A\mathcal{A}(F_t) - \delta - \rho\right) - \left(A\mathcal{A}(F_t) - x_t - \frac{C_F(N_t, e_t, F_t)}{K_t} - \delta\right)\right]. \tag{26}$$

Dynamics of F_t. Using Equation (13) and $\mathcal{S}_t/Y_t = 1 - C_t/Y_t = 1 - \frac{x_t}{A\mathcal{A}(F_t)}$, we obtain:

$$\dot{F}_t = \xi\left(\frac{N_t}{N_t + \bar{N}}\right)^{\omega}\left(1 - \frac{x_t}{A\mathcal{A}(F_t)}\right)^{\psi} F_t\left(1 - \frac{F_t}{F^{\max}}\right) - \delta_F F_t. \tag{27}$$

Finally, N_t is pinned down by the entry condition (16), where revenue depends on (F_t, x_t, K_t) via Equation (14) and costs depend on (N_t, e_t, F_t) via Equation (11), with effort $e_t = e^{\star}(\mathcal{S}_t, N_t)$ from Equation (19). This is the source of strong nonlinearity: N_t solves a fixed-point problem.

The model delivers multiple steady states when the implied mapping $N = \mathcal{N}(F, x, K)$ is sufficiently nonlinear. Intuitively: i) when F is low, $\mathcal{A}(F)$ is low, so Y/K is low; the economy saves less in equilibrium, and the fee base for finance is small. Few intermediaries enter, congestion is high, and F fails to accumulate—a low-finance trap; ii) when F is high, Y/K is high; the economy can sustain larger intermediation scale. Entry expands N, congestion falls, financial deepening accelerates, and high F is self-sustaining.

To state this cleanly, define a stationary pair (x^*, F^*) such that $\dot{x} = \dot{F} = 0$ and N satisfies Equation (16).[2]

Proposition 1 (Multiple steady states)**.** *Suppose $\mathcal{A}(F)$ is given by Equation (8), financial*

[2]Given the AK structure, levels of K scale out along balanced paths, and the relevant stationary objects are ratios and F.

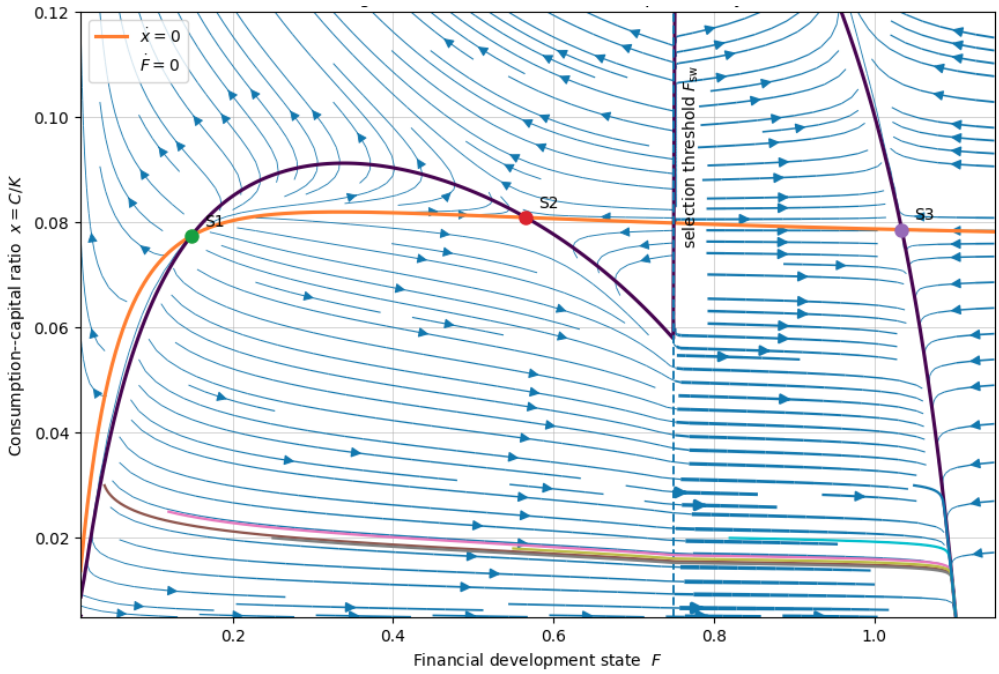
deepening follows Equation (13), and intermediation costs satisfy Equation (11)–(18). If: i) fixed operating cost f is large enough, ii) congestion curvature ν is large enough, and iii) deepening is sufficiently state-dependent (large enough ψ and/or $F^{\max}$ not too small), then there exist parameter values for which the stationary system admits three stationary points (x^*, F^*): a low-F stationary point, a middle threshold stationary point, and a high-F stationary point. The low and high stationary points are locally stable, while the middle stationary point is unstable.

Proof. See Appendix A. □

2.4 Phase diagram

Figure 1 depicts the phase diagram in the two-dimensional state space (F, x), where F is the stock of financial development and $x \equiv C/K$ is the consumption–capital ratio. The arrows show the direction of motion implied by the joint dynamics $(\dot{F}, \dot{x})$, while the thick curves are the two nullclines: the $\dot{x} = 0$ locus (consumption-growth balance) and the $\dot{F} = 0$ locus (financial deepening balance). Their intersections constitute steady states.

Figure 1: Phase diagram in $(F, x) = C/K$ with multiple steady states



The $\dot{x} = 0$ locus is obtained from the household Euler condition together with the capital-accumulation identity. Intuitively, along $\dot{x} = 0$ the consumption–capital ratio is exactly consistent with the net return on capital: if (F, x) lies above that locus, the model implies $\dot{x} < 0$ (consumption is high relative to capital, so x falls); if (F, x) lies below it, then $\dot{x} > 0$ (consumption is low relative to capital, so x rises). The $\dot{F} = 0$ locus collects points at which the forces

that expand the financial system (learning-by-intermediation and entry/network effects) exactly offset institutional depreciation.

In the plotted parameterization, the two nullclines intersect at three points: $S_1 : (F, x) = (0.147995, 0.077377), S_2 : (F, x) = (0.565835, 0.080908), S_3 : (F, x) = (1.034470, 0.078618)$. The vector field around these intersections shows a standard pattern of multiple steady states: S_1 and S_3 behave as locally stable long-run outcomes, whereas S_2 acts as a threshold (an unstable steady state). In particular, trajectories starting with initial conditions to the left of the separatrix implied by S_2 drift toward the low-finance steady state S_1, while initial conditions sufficiently far to the right converge toward the high-finance steady state S_3.

The existence of S_1 reflects a low-finance trap. When F is low, finance-augmented allocative efficiency is weak, the effective return to accumulating capital is limited, and the scale of intermediated saving is small. Entry is therefore muted, congestion is relatively severe, and the feedback from intermediation activity to financial deepening remains too weak to raise F persistently. The economy settles near a low F level where deepening stalls.

By contrast, S_3 is a high-finance regime. Once F is sufficiently high, the economy generates a larger intermediation base, entry expands, congestion costs are diluted, and the deepening process becomes self-reinforcing. Financial development then remains high and the economy converges to the steady state with large F.

The middle steady state S_2 is best interpreted as a critical mass condition. Near S_2, small adverse perturbations (e.g., a temporary disruption to the financial system) push the economy back toward S_1, while small favorable perturbations (e.g., a coordinated expansion in intermediation capacity) push it toward S_3. This provides a sharp mechanism for why economies with similar fundamentals can display persistent divergence in finance and macroeconomic performance.

The phase portrait also highlights that history matters: the same structural parameters can generate different long-run outcomes depending on initial conditions. The key reason is the nonlinearity embedded in the financial sector: fixed operating costs and congestion imply that, at low levels of activity, the per-unit cost of intermediation is high and entry is unattractive; at higher activity levels, entry becomes viable, congestion falls, and the marginal contribution of intermediation to deepening rises. This generates the characteristic S-shaped deepening dynamics and the triple intersection of $\dot{x} = 0$ and $\dot{F} = 0$.

The diagram suggests two empirically relevant implications. First, finance–growth data may exhibit threshold effects: marginal improvements in finance can have small effects below the critical region, but large effects once the economy is near or beyond the middle steady state. Second, policy interventions that shift the $\dot{F} = 0$ locus downward/upward—for instance by reducing operating costs of intermediation, improving enforcement, or lowering congestion—

can eliminate the low-finance trap by removing the middle intersection, thereby making the high-finance steady state globally attractive. In this sense, policies that expand intermediation capacity and improve efficiency are most potent when they help the economy cross the critical mass region around S_2.

3 Conclusion

This paper studies how financial development and economic growth can become jointly self-reinforcing. We build a parsimonious continuous-time general-equilibrium framework with households, producers, and financial intermediaries in which financial development is an endogenous stock. Finance raises allocative efficiency and therefore the effective return to capital accumulation, but intermediation also absorbs real resources through fixed operating costs and congestion-type losses. The central implication is a threshold mechanism: when the financial system is small, per-unit intermediation costs are high and deepening is weak, so the economy can be trapped in a low-finance, low-efficiency regime; once a critical mass is reached, entry and learning-by-intermediation reduce congestion and accelerate deepening, generating a high-finance, high-efficiency regime. The resulting phase diagram delivers multiple steady states in a transparent way and highlights why similar economies can diverge for long periods.

Two natural extensions are left for future work. One is to introduce default risk and endogenous crises to study whether the high-finance steady state is also more fragile. Another is to take the model to the data by estimating deepening and congestion parameters using micro-level intermediation cost measures and examining whether the implied threshold patterns match cross-country and within-country evidence.

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2005). Institutions as the fundamental cause of long-run growth. In P. Aghion & S. N. Durlauf (Eds.), *Handbook of Economic Growth* (Vol. 1A, pp. 385–472). Elsevier.

Arcand, J.-L., Berkes, E., & Panizza, U. (2015). Too much finance? *Journal of Economic Growth*, 20(2), 105–148.

Beck, T., Levine, R., & Loayza, N. (2000). Finance and the sources of growth. *Journal of Financial Economics*, 58(1–2), 261–300.

Bencivenga, V. R., & Smith, B. D. (1991). Financial intermediation and endogenous growth. *Review of Economic Studies*, 58(2), 195–209.

Cecchetti, S. G., & Kharroubi, E. (2012). Reassessing the impact of finance on growth. BIS Working Papers, No. 381, Bank for International Settlements.

Goldsmith, R. W. (1969). *Financial Structure and Development*. Yale University Press.

Greenwood, J., & Jovanovic, B. (1990). Financial development, growth, and the distribution of income. *Journal of Political Economy*, 98(5), 1076–1107.

King, R. G., & Levine, R. (1993). Finance and growth: Schumpeter might be right. *Quarterly Journal of Economics*, 108(3), 717–737.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1998). Law and finance. *Journal of Political Economy*, 106(6), 1113–1155.

Levine, R. (1997). Financial development and economic growth: Views and agenda. *Journal of Economic Literature*, 35(2), 688–726.

McKinnon, R. I. (1973). *Money and Capital in Economic Development*. Brookings Institution.

Rajan, R. G., & Zingales, L. (1998). Financial dependence and growth. *American Economic Review*, 88(3), 559–586.

Shaw, E. S. (1973). *Financial Deepening in Economic Development*. Oxford University Press.

Appendix

A Proof of Proposition 1

Step 0 (reduced 2D system): Let $c_t \equiv C_t/K_t$ and $y(F_t) \equiv Y_t/K_t = A\mathcal{A}(F_t)$. Under Equation (8),

$$y(F) \;=\; A\left(\frac{F}{\bar{F}+F}\right)^{\vartheta}, \qquad y'(F)>0, \quad \lim_{F\downarrow 0} y(F)=0, \quad \lim_{F\uparrow\infty} y(F)=A. \tag{A.1}$$

Define the (per unit of capital) saving flow:

$$s(F,c) \;\equiv\; \frac{Y-C}{K} \;=\; y(F)-c, \qquad \text{and the saving rate} \quad \sigma(F,c)\equiv\frac{s(F,c)}{y(F)} = 1-\frac{c}{y(F)}. \tag{A.2}$$

Under free entry, aggregate intermediation revenue equals aggregate intermediation resource costs. A convenient reduced-form implication of Equation (14)–(16) is that the resource cost absorbed by the financial system is proportional to the intermediated scale $Y-C$:

$$\frac{C_F}{K} \;=\; m(F)\,s(F,c), \qquad m(F)\equiv \mu\,\mathcal{B}(F) \;=\; \mu\left(\frac{F}{\bar{F}+F}\right)^{\eta}, \tag{A.3}$$

with $m(F)\in(0,\mu)$, $m'(F)>0$, and $\lim_{F\downarrow 0} m(F)=0$. This equality is the accounting identity implied by Equation (16): total revenue $\mu\mathcal{B}(F)(Y-C)$ is dissipated into the operating/screening/congestion costs C_F.

The household Euler equation and the capital accumulation identity yield the 2D system in (F_t, c_t):

$$\dot{c} = c\left[\frac{1}{\gamma}\big(y(F)-\delta-\rho\big) - \Big(y(F)-c-m(F)(y(F)-c)-\delta\Big)\right], \tag{A.4}$$

$$\dot{F} = F\left[\xi\, G(N)\,\sigma(F,c)^{\psi}\left(1-\frac{F}{F^{\max}}\right)-\delta_F\right], \qquad G(N)\equiv\left(\frac{N}{N+\bar{N}}\right)^{\omega}\in[0,1). \tag{A.5}$$

The key nonlinearity is that N is pinned down endogenously by the entry condition, and (crucially) can be multi-valued for given (F,c) when congestion is sufficiently curved.

Step 1 (the $\dot{c}=0$ locus is a graph $c=\bar{c}(F)$): Fix $F>0$. Define the bracket term in Equation (A.4) as $\Phi(F,c)$ so that $\dot{c} = c\,\Phi(F,c)$. On the economically relevant region $c>0$ and $s(F,c)=y(F)-c>0$, the equation $\dot{c}=0$ is equivalent to $\Phi(F,c)=0$, i.e.

$$\frac{1}{\gamma}\big(y(F)-\delta-\rho\big) = y(F)-c-m(F)\big(y(F)-c\big)-\delta = (1-m(F))\big(y(F)-c\big)-\delta. \tag{A.6}$$

Solving Equation (A.6) yields a unique $c = \bar{c}(F)$:

$$\bar{c}(F) = y(F) - \frac{\delta + \frac{1}{\gamma}\big(y(F) - \delta - \rho\big)}{1 - m(F)}. \tag{A.7}$$

Because $y(F)$ and $m(F)$ are C^1 and $m(F) < 1$ (choose $\mu < 1$), $\bar{c}(F)$ is continuous (indeed C^1) on any compact subset of $(0, \infty)$ where the interior condition $0 < \bar{c}(F) < y(F)$ holds. Hence the stationary system can be reduced to a one-dimensional condition on F along $c = \bar{c}(F)$.

Step 2 (two-entry equilibria: the entry equation can have two positive roots in N): Suppress effort for a moment; we return to it at the end of this step. Under Equation (11) with congestion $\Phi(x) = x^\nu$, the entry condition (16) at given (F, c) can be written (in per-unit-of-capital terms) as:

$$m(F)\, s(F,c) = fN + \chi \left(\frac{s(F,c)}{N + \underline{N}}\right)^\nu s(F,c). \tag{A.8}$$

Let $S \equiv s(F,c) > 0$ and define:

$$H(N; F, S) \equiv fN + \chi\, S^{\nu+1} (N + \underline{N})^{-\nu} - m(F)S. \tag{A.9}$$

Then Equation (A.8) is $H(N; F, S) = 0$ for $N \geq 0$. Note that:

$$H(0; F, S) = \chi\, S^{\nu+1} \underline{N}^{-\nu} - m(F)S, \tag{A.10}$$

$$\lim_{N \to \infty} H(N; F, S) = +\infty \quad \text{(because } f > 0\text{)}. \tag{A.11}$$

Moreover,

$$\frac{\partial H}{\partial N} = f - \chi\nu\, S^{\nu+1} (N + \underline{N})^{-(\nu+1)}. \tag{A.12}$$

Thus $\partial H/\partial N \to -\infty$ as $N \downarrow 0$ when ν is large and $S/\underline{N}$ is not too small, while $\partial H/\partial N \to f > 0$ as $N \to \infty$. Hence $H(\cdot; F, S)$ has a unique critical point (a unique global minimum) at $N = N^\star(F, S)$ satisfying $\partial H/\partial N = 0$:

$$N^\star + \underline{N} = S \left(\frac{\chi\nu}{f}\right)^{\frac{1}{\nu+1}}. \tag{A.13}$$

Evaluating H at $N^\star$ gives, after algebra,[3]

$$\frac{H(N^\star; F, S)}{S} = \underbrace{(\nu + 1)\left(\frac{f^\nu \chi}{\nu^\nu}\right)^{\frac{1}{\nu+1}}}_{\equiv\, \tilde{m}(\nu, f, \chi)} - m(F) - \frac{f\underline{N}}{S}. \tag{A.14}$$

[3] Substitute Equation (A.13) into Equation (A.9) and simplify using the identity $f\alpha + \chi\alpha^{-\nu} = (\nu + 1)(f^\nu\chi/\nu^\nu)^{1/(\nu+1)}$ for $\alpha = (\chi\nu/f)^{1/(\nu+1)}$.

Now fix $(\nu, f, \chi, \underline{N})$ and consider S not too small so that $f\underline{N}/S$ is negligible. If

$$m(F) \;>\; \bar{m}(\nu, f, \chi), \tag{A.15}$$

then $H(N^\star; F, S) < 0$ for sufficiently large S, while $H(0; F, S)$ can be made positive by taking ν large (so the congestion term $\chi S^{\nu+1}\underline{N}^{-\nu}$ explodes at $N = 0$ when $S/\underline{N} > 1$). Together with $H(N) \to \infty$ as $N \to \infty$, the intermediate value theorem implies that $H(\cdot; F, S) = 0$ has two positive roots, say

$$0 < N_L(F, S) < N^\star(F, S) < N_H(F, S). \tag{A.16}$$

Hence for a nonempty region of (F, S), the entry condition admits two distinct equilibrium levels of intermediary activity, a low-entry and a high-entry equilibrium. The region is nonempty because $m(F)$ is continuous and increasing in F, with $m(0) = 0$ and $\lim_{F\to\infty} m(F) = \mu$; choosing parameters so that $\mu > \bar{m}(\nu, f, \chi)$ ensures that (A.15) holds for all sufficiently large F.

Role of effort: Under Equation (18), congestion is multiplied by $\exp(-\zeta e)$ where e is chosen by intermediaries (or by a regulator) via a strictly convex trade-off as in Equation (19). This simply replaces χ by an effective $\tilde{\chi}(F, S, N) \equiv \chi \exp(-\zeta e^\star(F, S, N))$ with $\tilde{\chi} \in (0, \chi]$ and continuous in (F, S, N). The shape properties above (unique interior minimum and the possibility of two roots) are preserved, because $H(N)$ remains the sum of a linear term fN and a decreasing-in-N congestion term with sufficiently steep curvature when ν is large.

Step 3 (reducing the steady-state problem to a scalar equation): A stationary point (F^*, c^*) satisfies $\dot{c} = \dot{F} = 0$ with $F^* > 0$ and $c^* > 0$. From Step 1, $c^* = \bar{c}(F^*)$, and therefore the saving rate at a stationary point is:

$$\bar{\sigma}(F) \;\equiv\; \sigma(F, \bar{c}(F)) \;=\; 1 - \frac{\bar{c}(F)}{y(F)} \;=\; \frac{y(F) - \bar{c}(F)}{y(F)}. \tag{A.17}$$

Plugging $c = \bar{c}(F)$ into Equation (A.5) shows that stationary F^* must satisfy:

$$\Psi(F; N) \;\equiv\; \xi\, G(N)\, \bar{\sigma}(F)^\psi \Big(1 - \frac{F}{F^{\max}}\Big) - \delta_F \;=\; 0, \tag{A.18}$$

where N must also solve the entry condition (A.8) evaluated at $(F, \bar{c}(F))$. Because Step 2 implies that for some $(F, \bar{\sigma}(F))$ there exist two feasible N values, (A.18) can hold at different F's under different entry branches.

Step 4 (existence of three stationary points): Define the two branch functions on any interval where two entry equilibria exist:

$$N_L(F) \equiv N_L\big(F,\, S(F)\big), \qquad N_H(F) \equiv N_H\big(F,\, S(F)\big), \qquad S(F) \equiv y(F) - \bar{c}(F). \tag{A.19}$$

By the implicit function theorem (since $\partial H/\partial N \neq 0$ at simple roots), $N_L(F)$ and $N_H(F)$ are continuous on that interval.

Consider the scalar drift functions associated with each entry branch:

$$\Psi_L(F) \equiv \Psi\big(F; N_L(F)\big), \qquad \Psi_H(F) \equiv \Psi\big(F; N_H(F)\big). \tag{A.20}$$

We now show that one can choose parameter values (consistent with the proposition's qualitative restrictions) such that Ψ_L has two zeros and Ψ_H has one zero, implying three stationary points in total.

(i) Behavior at low F. For F sufficiently small, $m(F)$ is close to zero, so intermediation revenue is too small to cover fixed costs; the only feasible entry outcome is $N = 0$ (no active intermediaries). Then $G(N) = 0$ and Equation (A.5) implies:

$$\dot{F} \approx -\delta_F F < 0, \qquad \Rightarrow \qquad \Psi(F; 0) = -\delta_F < 0. \tag{A.21}$$

Hence the deepening drift is negative at low F.

(ii) Behavior near the upper bound $F^{\max}$. For any bounded N and any $\bar{\sigma}(F) \in [0, 1]$,

$$\lim_{F \uparrow F^{\max}} \Psi(F; N) = -\delta_F < 0, \tag{A.22}$$

because $\big(1 - F/F^{\max}\big) \to 0$. Thus the deepening drift is negative sufficiently close to $F^{\max}$.

(iii) Creating an interior region where $\Psi > 0$. Since $G(N)$ is increasing in N and $N_H(F) > N_L(F)$ whenever both exist,

$$0 \leq G\big(N_L(F)\big) < G\big(N_H(F)\big) < 1. \tag{A.23}$$

Choose ν large so that the entry equation admits two roots on a nondegenerate interval of F (Step 2), and choose ψ and ξ large enough so that for some interior $F = \tilde{F}$ on the high-entry branch,

$$\Psi_H(\tilde{F}) = \xi\, G\big(N_H(\tilde{F})\big)\, \bar{\sigma}(\tilde{F})^{\psi} \left(1 - \frac{\tilde{F}}{F^{\max}}\right) - \delta_F \;>\; 0. \tag{A.24}$$

This is always feasible because the first term in Equation (A.24) scales linearly in ξ and can be made arbitrarily large by increasing ξ (holding other parameters fixed), while remaining consistent with a bounded $G(\cdot)$ and $\bar{\sigma}(\cdot) \in (0, 1)$.

(iv) Three crossings. Combine Equations (A.21), (A.22), and (A.24). By continuity of $\Psi_H(F)$ on the interval where the high-entry root exists, the intermediate value theorem implies that $\Psi_H(F) = 0$ has at least one solution F_3^* in $(\tilde{F}, F^{\max})$. This delivers a high-finance stationary

point $(F_3^*, \bar{c}(F_3^*))$.

Next, because $N_L(F)$ is strictly smaller, the term $\xi G(N_L(F))\bar{\sigma}(F)^{\psi}(1 - F/F^{\max})$ can be made to form a smaller hump as a function of F, while still being positive in an intermediate region (by state dependence ψ and saturation $F^{\max}$). Concretely, choose parameters such that there exist $0 < F_1 < F_2 < F^{\max}$ (in the low-entry existence region) with

$$\Psi_L(F_1) > 0, \qquad \Psi_L(F_2) < 0, \tag{A.25}$$

while $\Psi_L(F)$ remains negative for F sufficiently close to 0 (by Equation (A.21) in the no-entry region). Then continuity implies there exist two distinct solutions $F_1^* \in (0, F_1)$ and $F_2^* \in (F_1, F_2)$ such that $\Psi_L(F_1^*) = \Psi_L(F_2^*) = 0$. These yield a low-finance stationary point $(F_1^*, \bar{c}(F_1^*))$ and an intermediate stationary point $(F_2^*, \bar{c}(F_2^*))$. Together with the high-entry root F_3^*, we obtain three stationary points.[4]

Step 5 (local stability classification): Linearize the planar system (A.4)–(A.5) around a stationary point (F^*, c^*), with $c^* = \bar{c}(F^*)$ and $\Psi(F^*; N^*) = 0$. Let J^* denote the Jacobian:

$$J^* = \begin{pmatrix} \partial_F \dot{F} & \partial_c \dot{F} \\ \partial_F \dot{c} & \partial_c \dot{c} \end{pmatrix}_{(F^*, c^*)}. \tag{A.26}$$

At any interior stationary point ($F^* > 0$, $c^* > 0$, $c^* < y(F^*)$), the partial derivatives satisfy the sign pattern:

$$\partial_c \dot{F} < 0, \qquad \partial_F \dot{c} > 0, \qquad \partial_c \dot{c} > 0. \tag{A.27}$$

The first inequality follows because $\dot{F}$ is increasing in the saving rate $\sigma = 1 - c/y$, so increasing c lowers σ and reduces deepening. The second follows because higher F increases $y(F)$ and therefore raises the net return environment in Equation (A.4). The third follows directly from Equation (A.4) because, holding F fixed, raising c reduces saving and raises c further (the ratio $c = C/K$ is a jump/control variable; saddle-path stability is the economically relevant notion).

Now note that at a stationary point, $\dot{F} = F \cdot \Psi(F; N)$ with $\Psi(F^*; N^*) = 0$, so

$$\partial_F \dot{F}\Big|_* = F^* \cdot \partial_F \Psi(F; N)\Big|_*, \qquad \partial_c \dot{F}\Big|_* = F^* \cdot \partial_c \Psi(F; N)\Big|_* < 0. \tag{A.28}$$

Hence the sign of $\partial_F \dot{F}$ is governed by $\partial_F \Psi$. Because the middle stationary point is constructed as a threshold where the deepening drift crosses from increasing to decreasing regimes, we can

[4] Existence of parameter values satisfying Equation (A.25) is straightforward because $\bar{\sigma}(F)$ and $(1 - F/F^{\max})$ are continuous, $\bar{\sigma}(F)$ can be made sharply state-dependent through a large ψ, and the amplitude difference between branches is controlled by $G(N_L)$ versus $G(N_H)$, which becomes large when congestion curvature ν is large and fixed costs f are nontrivial.

choose parameters (large ν and ψ) so that

$$\partial_F \dot{F}\Big|_{S_1} < 0, \qquad \partial_F \dot{F}\Big|_{S_3} < 0, \qquad \partial_F \dot{F}\Big|_{S_2} > 0, \tag{A.29}$$

i.e., the deepening drift is locally mean-reverting at the low and high stationary points but locally self-reinforcing at the intermediate point.

Given Equations (A.27)–(A.29), one can ensure saddle-path stability at S_1 and S_3 as follows. The determinant of J^* is:

$$\det(J^*) = (\partial_F \dot{F})(\partial_c \dot{c}) - (\partial_c \dot{F})(\partial_F \dot{c}). \tag{A.30}$$

Because $(\partial_c \dot{F})(\partial_F \dot{c}) < 0$ by Equation (A.27), the second term in (A.30) is positive. Thus $\det(J^*)$ is negative (a saddle) whenever the negative product $(\partial_F \dot{F})(\partial_c \dot{c})$ dominates in magnitude:

$$|\partial_F \dot{F}|\, \partial_c \dot{c} \;>\; -(\partial_c \dot{F})(\partial_F \dot{c}). \tag{A.31}$$

Condition (A.31) is satisfied when the F-dynamics are sufficiently steep locally, which is exactly what large congestion curvature ν (making N and hence $G(N)$ highly nonlinear) and large state dependence ψ (amplifying the response to the saving rate) accomplish. Hence S_1 and S_3 can be made saddle-path stable.

At the intermediate point S_2, $\partial_F \dot{F} > 0$ by Equation (A.29). With $\partial_c \dot{c} > 0$, the trace is positive whenever $\partial_F \dot{F}$ is not too small, and $\det(J^*)$ can be made positive by making $\partial_F \dot{F}$ sufficiently large (again feasible under large ν and ψ). In that case both eigenvalues of J^* are positive and S_2 is an unstable node (a threshold). This establishes the stated stability classification in the economically relevant sense: the low and high stationary points admit locally stable manifolds (saddle-path stability), while the middle point is unstable.

Conclusion. Steps 1–4 show that there exist parameter values satisfying the proposition's qualitative restrictions under which the system admits three stationary points. Step 5 shows that, under the same restrictions (and by strengthening steepness through large ν and ψ), the low and high stationary points are saddle-path stable while the middle point is unstable. This completes the proof.

B Interpretation of Parameters

For reference, below is a compact interpretation of all parameters used in the model.

Households.

- ρ: subjective discount rate; higher ρ lowers saving incentives and reduces long-run consumption growth given returns.
- γ: CRRA curvature (inverse IES); higher γ makes consumption growth less responsive to changes in returns.
- δ: physical capital depreciation rate.

Production and finance-augmented efficiency.

- A: baseline productivity parameter; scales output per unit of effective capital.
- $\mathcal{A}(F)$: allocative-efficiency term increasing in financial development.
- $\bar{F}$: normalization that controls the "half-saturation" level of finance; when $F = \bar{F}$, the fraction $F/(\bar{F} + F)$ equals $1/2$.
- ϑ: curvature/elasticity of allocative efficiency with respect to F; larger ϑ makes output more sensitive to changes in F at low-to-intermediate levels.
- $F^{\max}$: upper bound on sustainable financial development (technological/institutional frontier).

Intermediation costs and congestion.

- f: fixed operating cost per intermediary (branches, compliance, platform fixed costs). Larger f makes entry harder and strengthens threshold effects.
- κ: marginal cost scale of screening/monitoring intensity. Larger κ makes effort more expensive.
- χ: scale of real misallocation/congestion losses in finance (resources wasted in poorly screened intermediation).
- ν: curvature of congestion; larger ν means congestion increases rapidly when per-intermediary load rises, which is a key force behind multiple entry equilibria.
- $\underline{N}$: baseline capacity term preventing singularity at $N = 0$; interpretable as minimal backbone infrastructure.
- ζ: effectiveness of effort in mitigating congestion losses; larger ζ means monitoring is more powerful at reducing misallocation.

Financial deepening (law of motion for F).

- ξ: speed of financial deepening; higher ξ means a given level of activity/entry translates faster into institutional and technological improvements.
- $\bar{N}$: saturation parameter for the entry/network channel in Equation (13); larger $\bar{N}$ implies diminishing returns to additional intermediaries kick in later.
- ω: elasticity of deepening with respect to network size/entry; higher ω strengthens feedback from entry to F.
- ψ: elasticity of deepening with respect to intermediation intensity (saving rate). Higher ψ makes F more sensitive to changes in C/Y and therefore strengthens nonlinearity.
- δ_F: depreciation/obsolescence rate of financial development (institutional decay, technological obsolescence, erosion of trust).

Entry revenue side (reduced form).

- μ: revenue scale of intermediation (fee base per unit of intermediated saving). Larger μ makes entry easier.
- $\mathcal{B}(F)$: feasibility/fee-base term increasing in F (better enforcement and information increase scalable intermediation revenue).
- η: curvature of $\mathcal{B}(F)$; larger η implies revenues increase more sharply with F when F is low.

C Remarks for implementation and empirical mapping

Link to standard finance measures. In applications, F_t can be mapped to observed proxies such as private credit-to-GDP, payment-system penetration, or a composite financial development index. The model's $\dot{F}$ equation (13) suggests that deepening is faster when: i) i) the financial system is actively used (high saving/intermediation intensity) and ii) the intermediary network is sufficiently dense.

Why multiplicity is empirically plausible. The coexistence of fixed costs (fN) and congestion ($\mathcal{S}^{1+\nu}/(N+\underline{N})^{\nu}$) captures a common pattern: when the system is small, intermediation is expensive and error-prone; once scale and entry expand, per-unit costs fall and reliability

improves, which accelerates deepening. This is a natural source of threshold effects and multiple equilibria in finance-growth data.