Essays in Macro and Monetary Economics

By

Kui Huang

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Economics)

at the

University of Wisconsin-Madison

2016

Date of final oral examination: 07/11/2016

The dissertation is approved by the following members of the Final Oral Committee:

Randall Wright, Professor of Economics, Chair

Dean Corbae, Professor of Economics

Enghin Atalay, Assistant Professor of Economics

Erwan Quintin, Associate Professor of School of Business – Real Estate

Briana Chang, Assistant Professor of School of Business – Finance

Acknowledgments

I have been very lucky to receive aid and support from many people throughout my education at the University of Wisconsin-Madison. I owe my deepest gratitude to my advisor, Randall Wright. He has devoted countless hours to advising me. I am also grateful to Dean Corbae, Enghin Atalay, Erwan Quintin and Briana Chang who have been ideal members of my committee. They have always given me insightful comments and feedback on my research. Many other professors provided direction during my graduate school, Chao Gu, Cyril Monnet, Alberto Trejos, Rasmus Lentz, Kenneth West, Ananth Seshadri, Chao Fu, William Sandholm, Kenneth Hendricks and Jack Porter. I have benefited a lot from discussing with the students in the department, Han Han, Chao He, Yu Zhu and Michael Choi. Special thanks must be given to Zhewen Xu since the second part of this dissertation is joint work with him. Finally, I would like to thank my parents for continuous encouragement.

Contents

Abstract

This dissertation consists of two self-contained essays in macro and monetary economics, organized in the form of two chapters.

In the first chapter, I develop a model with limited commitment and endogenous monitoring to study the optimal number and size of banks. Banking arises endogenously because of economies of scale. The planner designates a fraction of ex-ante homogenous agents to be bankers and concentrates monitoring efforts on them. Having fewer bankers reduces total monitoring costs, but this means more deposits per banker. Having more deposits, however, increases the bankers' incentives to divert deposits for their own profit. The result is that the planner needs to give bankers some reward to dissuade such opportunistic behavior. The optimal number of banks is negatively related to the fixed and marginal monitoring costs, impatience, and the temptation to default, but positively related to the return on real investments. To implement efficient allocations, there is a tension between equilibrium with free entry and having positive bank profit for incentive reasons. When the tax on banks is not too high, there exist non-degenerate stationary equilibriums. The equilibrium allocation is optimal only if the government limits entry of banks. One natural way is to charge a tax on bankers and give a transfer to non-bankers; another way is to simply impose a quota by limiting the number of bank charters.

In the second chapter, using an overlapping generations model, I propose a resolution of the high household saving puzzle in China by analyzing the impact of the one-child policy and the resulting flattening of age-earning profiles on household saving behavior. Following Ben-Porath's (1967) human capital accumulation technology, with the implementation of the one-child policy, the initial human capital of each young worker who enters into the job market increases, which results in a decrease of the worker's on-the-job-training, and thus a flattening of age-earning profiles. The flattened age-earning profiles encourage younger cohorts to save more for consumption smoothing, and, therefore, provides an explanation for the high saving rates among the young. Both the data and the model demonstrate that the mechanism is valid.

Chapter 1

On the Number and Size of Banks: Efficiency and Equilibrium

1 Motivation

In the United States, between 1960 and 2014, the number of banks fell by more than half from about 13,000 to around 5,500. Between 1992 and 2014, the market share of the 10 largest banks grew dramatically from 21% to 57%. A great many of these changes started during the deregulation of bank size in the 1980s and 1990s. In 1960, banks could not branch across states and some states even forbade branching within a state. These legal and regulatory limits on bank size were subsequently removed. Figure 1 and figure 2 report the time paths for the number of banks and the market share of the 10 largest banks. I use two measures of bank size. The first is commercial bank assets, and the second is commercial bank deposits. I use fourth-quarter data on all commercial banks in the United States.[1]

My goal is to develop a theoretical model to address the following questions: Why did this structural change occur in the banking industry and is it desirable? Under what conditions is it socially optimal to have few large banks versus many small

[1] Following Berger, Kashyap, and Scalise (1995), I treat all banks and bank holding companies under a higher-level holding company as a single independent banking enterprise. For convenience, I will typically refer to each of these entities as a bank. Data on banks are taken from the Federal Deposit Insurance Corporation dataset.

banks? Why don't we want too few or too many banks? Is "unfettered competition" in banking optimal?

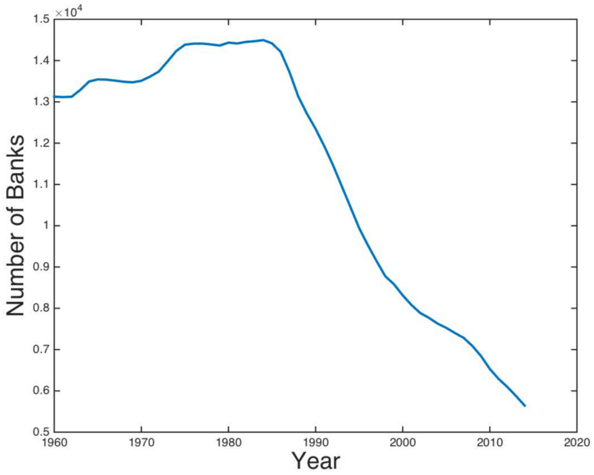

Figure 1: Drop in the number of banks

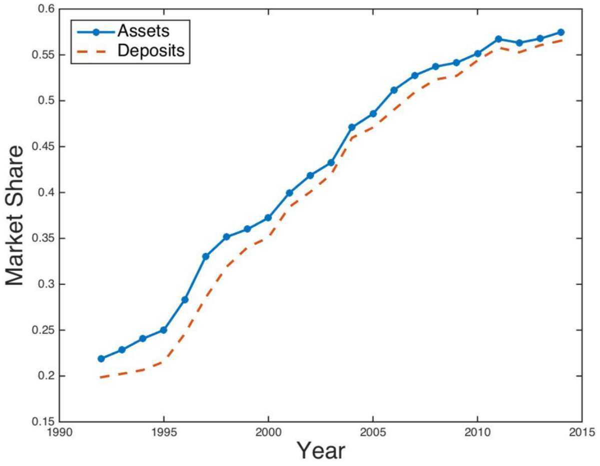

Figure 2: Market share of the 10 largest banks

I proceed with minimal assumptions about who bankers are or what they do. The agents that become bankers are ex-ante the same as the depositors. Obviously, some frictions are needed because models such as Arrow-Debreu have no roles for banks. There are two frictions in my model, arising from limited commitment: The agents that become bankers have a temptation to abscond with the proceeds (as in the cash-diversion models of Demarzo and Fishman 2007, or Biais et al. 2007), and there is imperfect monitoring (or imperfect record keeping). Responding to a classic challenge in monetary economics—what makes money essential—I want to first ask what makes banking arrangements essential.[2]

The formal model incorporates the following ingredients. Time is discrete and continues forever. Each period is divided into two subperiods. There are two types of infinitely lived agents: type 1 and type 2. Type 1 agents consume in the first subperiod while type 2 agents consume in the second. Both types produce the other type's consumption goods in the first subperiod. In a first-best world, it would be efficient to have type 2 deliver his production good to type 1 in the first subperiod, enabling type 1 to consume first, then invest and deliver his production good to type 2 in the second subperiod. In the second subperiod, however, type 1 is tempted to abscond with the proceeds. If type 1 defaults, type 2 knows it and needs to pay a monitoring cost to verify the default (communicate with the mechanism, or the court/legal system). One example of such costly communication is a lawsuit. With probability π, the mechanism records it, and the deviating type 1 is punished to future autarky. In general, we need to impose an incentive constraint guaranteeing

[2] I want to know which frictions lead to banking. As in Townsend (1988): "the theory should explain why markets sometimes exist and sometimes do not, so that economic organization falls out in the solution to the mechanism design problem" . Relatedly, I stick to a generalization of Wallace's (1998) dictum: "money should not be a primitive in monetary theory—in the same way that a firm should not be a primitive in industrial organization theory or a bond a primitive in finance." By extension, banks should not be a primitive in banking theory; they should arise endogenously.

type 1 does not default.

An efficient mechanism is to designate a fraction of the ex-ante homogenous type 1 to be bankers and concentrate monitoring efforts on them. A banker in my model is an agent that has three features: he takes deposits, makes investments on behalf of depositors, and his liabilities (claims on deposits) facilitate third-party transactions. It is because his activity resembles banking that I call him a banker. Of course, banks may do more, such as providing liquidity insurance or information processing. I downplay these functions, which have been studied extensively elsewhere, and focus instead on banking arising endogenously as a response to commitment problems and economies of scale.

Consider the cost-benefit trade-off of decreasing the number of bankers from the planner's perspective: Having fewer bankers reduces total monitoring costs, but this means more deposits per banker. Having more deposits, however, increases the bankers' incentives to divert deposits for their own profit, so that they may need to be monitored more rigorously. One important result is that the planner needs to give the bankers some reward to dissuade opportunistic behavior.

To implement efficient allocations in decentralized competitive markets, there is a tension between equilibrium with free entry and having positive bank profit for incentive reasons. Since bankers have higher payoffs than the non-bankers, all the agents would want to be bankers, which will lead to excess entry and thus an inefficient equilibrium. To increase welfare, the government needs to limit entry of banks, either by charging a tax on bankers or rationing bank charters. If the tax on banks is not too high, there exist stationary equilibriums with banks; if the tax on banks is higher than the cut-off value, there exists an equilibrium with no banks. For a given tax, we can have too much or too little entry, compared with the efficient outcome. If the tax is almost zero, nearly everyone wants to be a banker, and, thus, there is too much entry. If the tax is almost at the cut-off value, there is too little entry. When the government charges an optimal tax on the banker and gives an optimal transfer to the non-banker, the competitive equilibrium is efficient.

In the decentralization, inside money also helps to implement efficient outcomes. Specifically, type 1 non-banker deposits his production with the banker, who invests on his behalf. The bankers issue receipts for deposits to type 1 non-bankers, which are then transferred to type 2 in the first subperiod and redeemed in the second. The receipts, like bank notes through history, and later checks and debit cards, constitute a transactions medium—inside money.[3]

In the efficiency part, I derived the effects of parameter changes and thus can answer under what conditions it is socially optimal to have few large banks versus many small banks. The optimal number of banks is negatively related to the fixed and marginal monitoring costs, impatience, and the temptation to default, but positively related to the return on real investments. Second, it can explain why the the number of banks dropped in the United States. Because the world is more complex than before and it's easier for people to cheat, the temptation to default increases. According to the effects of parameter changes, with the rise of the temptation to

[3] This is a commonly understood role of banking. Consider Selgin (2007): “Genuine banks are distinguished from other kinds of financial intermediaries by the readily transferable or spendable nature of their IOUs, which allows those IOUs to serve as a means of exchange, that is, money. Commercial bank money today consists mainly of deposit balances that can be transferred either by means of paper orders known as checks or electronically using plastic debit cards.”

default, the number of banks decreases. Third, the model can explain why we do not want too few banks. The recent literature has stressed "financial fragility" or "too big to fail" as interpretations, but I propose a different explanation. If we have too few banks, each bank would have too many deposits and this increases their incentive to misbehave. It can also explain why we cannot have too many banks, because of the monitoring cost. Finally, it can explain whether "unfettered competition" in banking is optimal. Free competition will lead to too much entry compared with the efficient outcome.

The model is related to several papers about credit with limited commitment, such as Kehoe and Levine (1993), Alvarez and Jermann (2000), and Gu et al. (2013a), but the application and emphasis concern banking.

In terms of the mainstream banking literature, Gorton and Winton (2002) and Freixas and Rochet (2008) provide surveys. One approach, originated by Leland and Pyle (1977) and developed by Boyd and Prescott (1986), interprets banks as information-sharing coalitions. Another strand, pioneered by Diamond and Dybvig (1983), interprets banks as coalitions providing liquidity insurance. A related approach, following Diamond (1984) and Williamson (1986, 1987), interprets banks as delegated monitors taking advantage of returns to scale. I abstract from liquidity provision and information sharing, and instead highlight banking arising endogenously as a response to commitment problems and economies of scale. Compared with Diamond (1984) and Williamson (1986, 1987), a big advantage of my paper is that it is an infinite-horizon model.[4] It allows banker's reputation to have a role ("reputation"

[4] Diamond (1984) and Williamson (1986) are finite-horizon models, and Williamson (1987) is an overlapping generations model where each agent lives for two periods.

in the sense of Kehoe-Levine). Also, bankers have the incentive to honor their notes that circulate; this would not happen in a finite-horizon world because they would choose not to redeem the notes in the last period. Another major difference from most banking literature is that who is a banker, plus how many plus how big, are all endogenous variables.

I also highlight literature where bank liabilities are payment instruments, such as Gu et al. (2013b), Cavalcanti and Wallace (1999a, 1999b), and He et al. (2005, 2008). My model is based on but different from Gu et al. (2013b). In their paper, banking arises endogenously because of heterogeneity, some people are more trustworthy to be bankers.[5] More trustworthy agents accept deposits by less trustworthy agents and invest them. Then these less trustworthy agents use their claims on deposits to facilitate trade with third parties. While in this model, even if the bankers and depositors are ex-ante homogenous, banking can still arise because of economies of scale. Another difference is that the monitoring probability is exogenous in their paper, whereas I endogenize it. Compared with Cavalcanti and Wallace (1999a, 1999b), and He et al. (2005, 2008), where inside money also facilitates trade, a major difference is that they do not have deposits, delegated investments or endogenous monitoring.[6]

With regard to literature on bank number and bank size, there are some empirical

[5] In Gu et al. (2013b), agents are better suited to banking when they have a good combination of the following characteristics that make them more trustworthy: they are relatively patient; they are more visible, by which they mean more easily monitored; they have a greater connection to the economic system; they have access to better investment opportunities; and they derive lower payoffs from opportunistically diverting resources.

[6] In addition, see Wallace (2005), Koeppl et al. (2008), Andolfatto and Nosal (2009), Huangfu and Sun (2011), Mills (2008), Sanches andWilliamson (2010), and Monnet and Sanches (2012).

papers. Janicki and Prescott (2006) document the changes in the size distribution of U.S. banks between 1960 and 2005, but they don't provide a theory. Corbae and D'Erasmo (2013) is one of the few papers where both the number and size of banks are endogenously determined. However, their work focuses on the industrial organization approach to banking. They analyze a Stackelberg game between banks and the endogenous bank size distribution arises out of entry and exit in response to shocks to borrowers' production technologies. They focus on mechanisms such as "too big to fail", while I look at something else. Also, a main goal here is a tractable if somewhat stylized framework, so that it is possible to derive analytic and not only numerical results.

The other related literature is that of monitoring. Monitoring has a broad sense of meanings. In Diamond (1984), and Townsend (1979), it means punishing or auditing a borrower who fails to meet contractual obligations in the context of costly state verification. In Broecker (1990), it means screening projects a priori in the context of adverse selection. In Holmstrom and Tirole (1997), and Diamond and Rajan (2001), it means preventing opportunistic behavior of a borrower during the realization of a project (moral hazard). The monitoring in my paper is similar to Diamond (1984), in which the deviation is costly to verify. If there is a default, the banker is detected by the mechanism with probability π, and punished to future autarky with payoff 0.

The rest of the paper is organized as follows. Section 2 describes the basic environment without banking, which provides a simple model of credit with limited commitment and imperfect monitoring. Section 3 describes the environment with banking. Section 4 solves the planner's problem. It gives us a basic idea of how we can get the optimal number and size of bankers as well as incentive-feasible and efficient allocations using endogenous monitoring from the planner's perspective. All of the analysis here focuses on stationary allocations. Section 5 describes the decentralization, which shows how to implement efficient allocations using inside money (bank notes). Section 6 is Conclusion.

2 Environment without Banking

Time is discrete and continues forever. Each period is divided into two subperiods. There are two types of agents: measure 1 of type 1 agents, and measure 1 of type 2 agents. Type 1 agents consume good x and produce good y; type 2 agents consume good y and produce good x. Both goods are produced in the first subperiod; good x is consumed in the first subperiod, while good y is consumed in the second. There is a role for credit since type 1 consumes before type 2, and there is a notion of collateral since good y is produced in the first subperiod. Type 1 agents store and invest good y across subperiods, with fixed gross return ρ in terms of second-subperiod goods. There is no investment across periods, only across subperiods. This may be as simple as pure storage, perhaps for safekeeping, or any other investment; merely for ease of presentation do we impose a fixed return. To generate gains from trade in a simple way, type 2 agents cannot invest for themselves; more generally, we could let them invest, just not as efficiently. We can interpret type 1 agents as borrowers and type 2 agents as lenders.

Utility of type 1 is $U^1(x, y)$, and utility of type 2 is $U^2(\rho y, x)$. Both utility functions are strictly increasing in consumption and decreasing in production, strictly concave, twice differentiable, and $U^j(0,0) = 0$, $j = 1, 2$.

The timeline of the environment with credit is shown in Figure 3

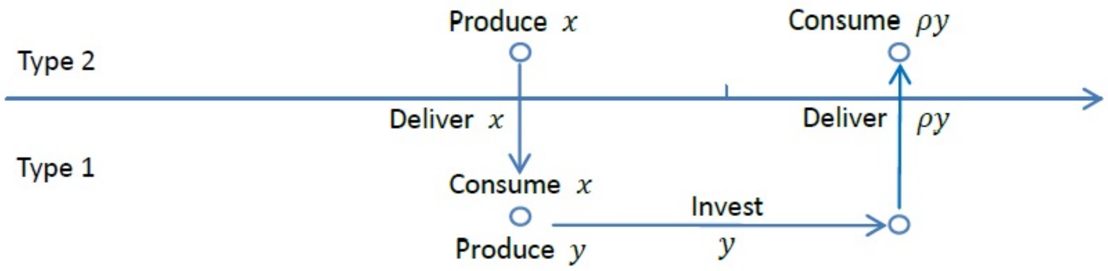

Figure 3: Timeline of the environment with no banking

There are two important frictions:

- Limited Commitment.

  When type 1 agents are supposed to deliver the goods, in the second subperiod, they can renege to obtain a payoff $\lambda\rho y$, over and above $U^1(x, y)$. This is the key incentive issue in the model. If $\lambda = 0$, investment constitutes perfect collateral, since type 1 agent has no gains from reneging when the production cost is sunk. However, if $\lambda > 0$, there is an opportunity cost to deliver the goods. Formally, diversion can be interpreted as type 1 agent consuming the investment returns, but it stands in for the more general idea that investors can divert resources opportunistically.

- Imperfect monitoring.

  Any deviation from the suggested outcome is detected by the mechanism with probability π, punished with future autarky with payoff 0,[7] and is not detected by the mechanism with probability $1-\pi$. Here, π is endogenous, which means the mechanism can choose monitoring intensity.

  We have many ways to rationalize this monitoring probability; a straightforward one is to assume imperfect record keeping: information concerning deviations "gets lost" with probability $1 - \pi$ across periods. More specifically, if a type 1 agent defaults, the type 2 agent who got defaulted on knows it and needs to verify the default (communicate with and report it to the mechanism, or court/legal system). One example of such costly communication is a lawsuit. With probability π, the mechanism (court/legal system) knows it and records it, and the deviator is punished to future autarky. There are various elements required to punish a deviation: (1) it must be observed by someone; (2) it must be communicated with the mechanism; and (3) it must be recorded/remembered. Failure on one of these dimensions—which is called imperfect memory by Kocherlakota (1998)—is enough to hinder punishments based on reputation.

[7] We can consider weaker punishments but this is obviously the most effective.

Assume monitoring with probability π implies a utility cost $c(\pi)$, where

$$c(\pi) = \begin{cases} k_0 + \pi k & \text{if } \pi > 0 \\ 0 & \text{if } \pi = 0 \end{cases} \tag{1}$$

The cost is paid by type 2 agent. Here, k_0 is a fixed cost, and k is a marginal cost, and the cost function implies increasing returns to scale (economies of scale).

The incentive feasible set with no commitment entails two participation constraints for type 1 and type 2 agents and one repayment constraint for type 1 agent. All of the analysis here focuses on stationary allocations.

$$U^1(x,y) \geq 0, \tag{2}$$

$$U^2(\rho y, x) - c(\pi) \geq 0, \tag{3}$$

$$U^1(x,y) + \beta V^1(x,y) \geq U^1(x,y) + \lambda \rho y + (1-\pi)\beta V^1(x,y), \tag{4}$$

where $V^1(x,y) = U^1(x,y)/(1-\beta)$ is the continuation value for the type 1 agent. When type 1 agent invests y, he promises to deliver ρy in the second subperiod, but he can always renege for a short-term gain $\lambda \rho y$, and so he delivers the goods only if the repayment constraint satisfies. The LHS is the payoff of not deviating, and the RHS is the payoff to behave opportunistically, again caught with probability π, and punished to future autarky with payoff 0. Note that $U^1(x,y)$ is sunk at the time of decision. The repayment constraint reduces to $U^1(x,y) \geq \frac{(1-\beta)\lambda\rho y}{\beta\pi} = \frac{r\lambda\rho y}{\pi}$ where $r = (1-\beta)/\beta$. A high r or high λ both increase the temptation to default. We say an agent is more trustworthy when he has smaller $r\lambda$, which means he can credibly promise more (or has better credit).[8]

3 Environment with Banking

The planner designates measure μ of type 1 agents to be bankers and concentrates monitoring efforts on them. The other measure $1-\mu$ of type 1 agents are non-bankers. (I will explain why those measure μ of agents resemble bankers and why their activity resembles banking later.) The type 1 bankers and non-bankers are ex-ante homogeneous. Each type 1 non-banker produces part y_n of good y, deposits his production with type 1 banker, and consumes part x_n of good x. Each type 1 banker produces part y_b of good y, accepts deposits from type 1 non-banker, and consumes part x_b of good x. The bankers can store and invest the combined good y, from their own production and the deposits from the non-bankers, across subperiods, with fixed gross return ρ in terms of second-subperiod goods. The size (assets) of each bank after investment is $\rho y/\mu$.

The cost-benefit trade-off is that having fewer bankers reduces total monitoring costs, but this means more deposits per bank. Having more deposits, however, increases the bankers' incentives to divert deposits for their own profit, and thereby reduces the benefit to the economy.

[8] In Gu et al. (2013b), they have one more parameter γ, which is the probability that an agent will want to participate in the "market" each period. This "attachment to the market" parameter provides one more way to make an agent more or less trustworthy, since agents more attached to the market can be more trustworthy. Because it operates very much like r or λ, I omit it.

There are two feasibility constraints for good x and good y. If the type 2 agent produces good x, each type 1 banker consumes part x_b of good x, and each type 1 non-banker consumes part x_n of good x, then $x = \mu x_b + (1-\mu)\, x_n$. Similarly, if each type 1 banker produces part y_b of good y and each type 1 non-banker produces part y_n of good y, we can define $y \equiv \mu y_b + (1-\mu)\, y_n$. Type 1 bankers store and invest y in total across subperiods, get ρy after investment, and deliver the goods to type 2. Each type 2 agent consumes good ρy.

Utility of type 1 banker is $U^1(x_b, y_b)$, utility of type 1 non-banker is $U^1(x_n, y_n)$, and utility of type 2 is $U^2(\rho y, x)$. Both utility functions are strictly increasing in consumption and decreasing in production, strictly concave, twice differentiable, and $U^j(0,0) = 0$, $j = 1, 2$. We assume a discount factor across periods $\beta \in (0,1)$, there is no discount across subperiods with no loss in generality.

The timeline of the environment with banking is shown in figure 4.

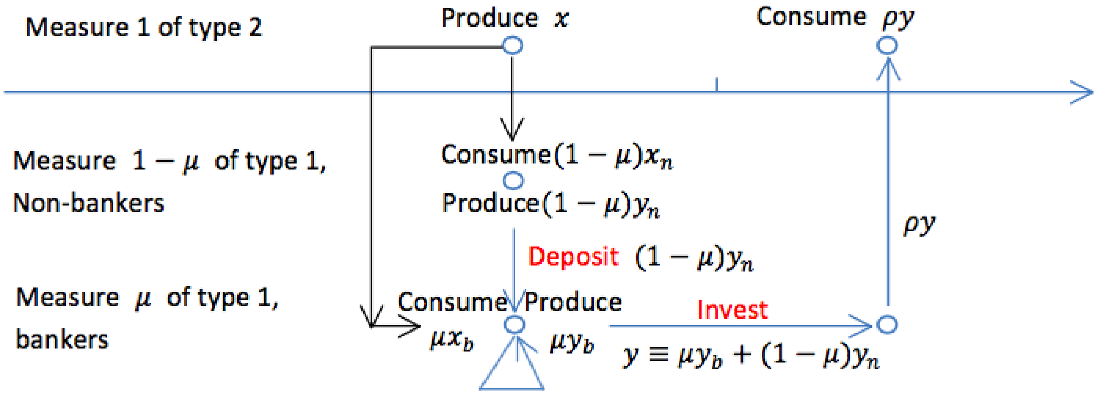

Figure 4: Timeline of the environment with banking

The banker in the model is an agent that has three features: he takes deposits, and makes investments on behalf of depositors; and his liabilities (claims on deposits) facilitate third-party transactions. The non-bankers here are depositors. I downplay other functions of banks, such as providing liquidity insurance or information processing, but these can be added using standard methods. Notice that a special case is $\mu = 1$; then we are back to the previous model with pure credit, where there are no depositors and, thus, no banking, all the type 1 agents can invest their own production goods and are tempted to divert the resources for their own profit. However, I am going to show that it is better for the planner to choose $\mu < 1$. Since we have a fixed monitoring cost k_0, it is better to monitor some of the people more intensely, and economize the number of bankers. Why would the planner not choose μ to be a tiny ϵ? In this case, each banker would have so many deposits, and they are more likely to default. Thus, the optimal number of banks is an interior solution. I define the case of no trade to be $\mu = 0$.

We will discuss both the planner's problem and the decentralization in the following two sections. In the planner's problem, the mechanism designer can recommend the agents how much to consume and how much to produce. While in the decentralized case, it shows how to implement the efficient allocations using inside money (bank note). One way to implement this is to have bankers issue receipts for deposits, which are then transferred to type 2 in the first subperiod and redeemed in the second. The receipts, like bank notes through history, and later checks and debit cards, constitute a transactions medium—inside money.

Here we can compare the theory with some facts from banking history. Institutions that accepted commodity deposits were operating long before the invention of coinage, let alone fiat currency. As Davies (2002) describes the situation, in ancient Mesopotamia and Egypt, goods were often deposited in temple and palace based banks, and, later, private banking houses. "Grain was the main form of deposits at first, but in the process of time other deposits were commonly taken: other crops, fruit, cattle and agricultural implements, leading eventually and more importantly to deposits of the precious metals. Receipts testifying to these deposits gradually led to transfers to the order not only of depositors but also to a third party." In ancient Babylon, also, as Ferguson (2008) says: "Debts were transferable, hence pay to the bearer rather than a named creditor. Clay receipts or drafts were issued to those who deposited grain or other commodities at royal palaces or temples." And, also as in the model, "the foundation on which all of this rested was the underlying credibility of a borrower's promise to repay."

4 Efficiency

Now what a planner or mechanism can do is to recommend an incentive feasible allocation in the group, as long as no one wants to deviate. All of the analysis here focuses on stationary allocations. The monitoring cost is paid by type 2, and the utility of type 2 is $U^2(\rho y, x) - \mu\,(k_0 + \pi k)$

We can define the ex post (conditional on type) welfare as

$$\begin{aligned} W(x_b, y_b, x_n, y_n, x, y) = \;& \theta\left[\mu U^1(x_b, y_b) + (1-\mu)\,U^1(x_n, y_n)\right] \\ & + (1-\theta)\left[U^2(\rho y, x) - \mu\,(k_0 + \pi k)\right] \end{aligned} \tag{5}$$

where we put the same weight θ on type 1 banker and non-banker, since they are ex-ante homogeneous, and weight $1-\theta$ on type 2 agent.[9]

The incentive feasible set with no commitment should satisfy the following participation constraints and incentive constraints:

Participation constraints for type 1 banker, non-banker and type 2 agent

$$U^1(x_b, y_b) \geq 0 \tag{6}$$

$$U^1(x_n, y_n) \geq 0 \tag{7}$$

$$U^2(\rho y, x) - \mu \left(k_0 + \pi k\right) \geq 0 \tag{8}$$

Repayment constraint for type 1 banker

$$U^1(x_b, y_b) + \beta V^1(x_b, y_b) \geq U^1(x_b, y_b) + \lambda \rho y / \mu + (1-\pi)\beta V^1(x_b, y_b) \tag{9}$$

where $V^1(x_b, y_b) = \frac{U^1(x_b, y_b)}{1-\beta}$ is the continuation value for type 1 banker. The LHS is the payoff from following the recommendation, while the RHS is the deviation payoff. This reduces to

$$U^1(x_b, y_b) \geq r\lambda\rho y / \pi\mu \tag{10}$$

where $r = \left(1-\beta\right)/\beta$. From expression (10), as we decrease the number of banks, the repayment constraint is tighter. If the number of banks is too small, this repayment constraint could be violated. On the other hand, from the welfare function, if the

[9] When we put the same weight on type 1 banker and non-banker, it's like there is a lottery where the planner randomly puts μ of type 1 agents as bankers and $1-\mu$ of them as non-bankers, and the summation of utility implies an ex-ante expected utility of a representative type 1 agent.

number of banks is too large, the monitoring cost would be too high. Thus, the optimal number of banks is interior.

To sum up, the incentive feasible set with no commitment should satisfy (7), (8) and (10) above. A planner can recommend an incentive feasible solution (x_b, y_b, x_n, y_n, x, y, π, μ) in the group.

$$\begin{aligned}
&\max_{(x_b,y_b,x_n,y_n,x,y,\pi,\mu)} \left\{\theta \left[\mu U^1(x_b,y_b) + (1-\mu)\, U^1(x_n,y_n)\right] \right.\\
&\qquad \left. + (1-\theta)\left[U^2(\rho y, x) - \mu\,(k_0 + \pi k)\right]\right\}\\
\text{s.t.}\quad & \mu x_b + (1-\mu)\, x_n = x\\
& \mu y_b + (1-\mu)\, y_n = y\\
& U^1(x_n,y_n) \geq 0\\
& U^2(\rho y, x) - \mu\,(k_0 + \pi k) \geq 0\\
& U^1(x_b,y_b) \geq \tfrac{r\lambda\rho y}{\pi\mu}
\end{aligned}$$

Lemma 1. *The repayment constraint must bind,* $U^1(x_b,y_b) = \frac{r\lambda\rho y}{\pi\mu}$.

Proof: If not, we could reduce π to increase the objective function. □

With a bit more structure on preferences, by using quasi-linearity as is usual in these models, we can get even more predictions, especially clean comparative statics results.

Suppose

$$\begin{aligned} U^1(x_b, y_b) &= u(x_b) - y_b \\ U^1(x_n, y_n) &= u(x_n) - y_n \\ U^2(\rho y, x) &= \rho y - v(x) \end{aligned}$$

where u is strictly increasing and concave, satisfies Inada conditions: $\lim_{x \to 0} u'(x) = +\infty$, $\lim_{x \to +\infty} u'(x) = 0$, and v is strictly increasing and convex.

From the binding repayment constraint for type 1 banker, $u(x_b) - y_b = \frac{r\lambda\rho y}{\pi\mu}$, we know $\pi = \frac{r\lambda\rho y}{[u(x_b)-y_b]\mu}$. Substituting $\pi = \frac{r\lambda\rho y}{[u(x_b)-y_b]\mu}$, $x = \mu x_b + (1-\mu)\, x_n$, and $y_n = \frac{y-\mu y_b}{1-\mu}$ into the planner's problem, it becomes

$$\begin{aligned} &\max_{(x_b, y_b, x_n, y, \mu)} \{\theta\, [\mu u(x_b) + (1-\mu)\, u(x_n) - y] \\ &\quad + (1-\theta) \left[\rho y - v\left(\mu x_b + (1-\mu)\, x_n\right) - \mu k_0 - \tfrac{r\lambda\rho y k}{u(x_b)-y_b}\right]\right\} \end{aligned}$$

FOCs

$$u'(x_b) - \tfrac{1-\theta}{\theta} v'(x) + \tfrac{1-\theta}{\theta} \tfrac{r\lambda\rho y k u'(x_b)}{\mu[u(x_b)-y_b]^2} = 0$$

$$y_b^* = 0$$

$$u'(x_n) - \tfrac{1-\theta}{\theta} v'(x) = 0$$

$$-1 + \tfrac{1-\theta}{\theta} \left[\rho - \tfrac{r\lambda\rho k}{[u(x_b)-y_b]}\right] = 0 \Rightarrow x_b^* = \overline{x_b}\,(k, r, \lambda, \rho)$$

$$u(x_b) - u(x_n) - \tfrac{1-\theta}{\theta} v'(x)\,(x_b - x_n) - \tfrac{1-\theta}{\theta} k_0 = 0$$

Proposition 1. *$x_b^* > x_n^*$, and $y_b^* = 0$. That is to say, the bankers can consume more* *than the non-bankers and do not need to produce.*[10]

Proof: See the Appendix.

The intuition is that the planner needs to give the bankers some reward to dissuade opportunistic behavior (satisfies the repayment constraint).

Proposition 2. $\frac{\partial x_b^*}{\partial k} > 0$, $\frac{\partial x_b^*}{\partial r} > 0$, $\frac{\partial x_b^*}{\partial \lambda} > 0$, *and* $\frac{\partial x_b^*}{\partial \rho} < 0$.

Proof: See the Appendix.

As the marginal monitoring cost increases, the first order effect is to reduce monitoring probability, and, thus, the bankers are more likely to renege, we have to compensate them more such that they don't deviate. Similarly, if the interest rate (impatience) increases, or if there is more temptation to behave badly, we need to give the bankers more compensation. If the rate of return increases, we can give the bankers less compensation.

Proposition 3. *Suppose* $u(x) = \frac{x^{1-\alpha}-1}{1-\alpha}$, *where* $\alpha > 0$, we have $\frac{\partial\left(x_b^*-x_n^*\right)}{\partial k} > 0$, $\frac{\partial\left(x_b^*-x_n^*\right)}{\partial r} > 0$, $\frac{\partial\left(x_b^*-x_n^*\right)}{\partial \lambda} > 0$, $\frac{\partial\left(x_b^*-x_n^*\right)}{\partial \rho} < 0$.

Proof: See the Appendix.

Note that $x_b^* - x_n^*$ is the premium that the bankers take because of the commitment problems. It's sort of the rent extracted by the banker. The premium that the bankers take is positively related to the marginal monitoring cost, the interest rate

[10] If we use the general additively separable utility, $U^1(x, y) = u(x) - v(y)$, we can get $x_b^* > x_n^*$ and $y_b^* < y_n^*$, the bankers can consume more than the non-bankers and produce less. With quasilinear utility, however, bankers specialize to just invest and not produce.

(impatience), and the temptation to default, but negatively related to the rate of return.

Proposition 4. $\frac{\partial \mu^*}{\partial k_0} < 0$, $\frac{\partial \mu^*}{\partial k} < 0$, $\frac{\partial \mu^*}{\partial r} < 0$, $\frac{\partial \mu^*}{\partial \lambda} < 0$, *and* $\frac{\partial \mu^*}{\partial \rho} > 0$.

Proof: See the Appendix.

As the fixed monitoring cost increases, we definitely should have fewer bankers. We can interpret the other comparative statics results of the optimal number of bankers through the premium that the bankers take. As the monitoring cost (or impatience, or the temptation to default) increases, we should have less bankers because it's more expensive to use them. As the rate of return increases, we should have more bankers because it's cheaper to use them.

The proposition can explain why the the number of banks dropped in the United States. Because the world is more complex than before and it's easier for people to cheat, the temptation to default λ increases. According to the effects of parameter changes, with the rise of the temptation to default, the number of banks decreases.

5 Equilibrium

From the planner's problem, we can get the second-best solution with frictions (limited commitment). Then I want to find a decentralized pricing mechanism such that the second-best allocations can be realized. Here, I am using the Walrasian pricing mechanism, where everyone takes prices as given. Because the agents who are selected to be bankers have a higher payoff than the non-bankers in the planner's problem, all the agents would want to be bankers, which is not efficient. Thus, the government needs to limit entry of banks. one natural way is to charge a tax τ on bankers and give a transfer t to non-bankers; another way is to simply impose a quota by limiting the number of bank charters.

5.1 Charging a Tax

To implement the efficient outcomes, we also need inside money (bank notes). When a type 1 non-banker wants to consume in the first subperiod, he produces and deposits output y_n with a type 1 banker in exchange for a receipt. Think of the receipt as a bearer note for goods y. He then gives this note to a type 2 agent in exchange for his consumption good x_n. Naturally, the type 2 agent accepts it, and carries this note to the second subperiod. Each type 1 banker borrows $\hat{y}$ from the non-banker, produces y_b by himself, and gives some notes to a type 2 agent in exchange for his consumption good x_b. When the type 2 agent wants to consume in the second subperiod, he redeems all the notes for his consumption good. Type 1 banker pays type 2 agent out of deposits—principal plus return on investments, ρy—to clear, or settle, the obligation. In this way the bank liabilities serve as inside money, like banknotes, checkbooks and debit cards. In sum, there are three types of trades. In the first subperiod, agents trade good x and bank notes issued by the banker; type 1 non-banker and banker trade good y and banknotes. In the second subperiod, A banknote entitles type 2 one unit of good y from the banker. The timeline is shown in Figure 5.

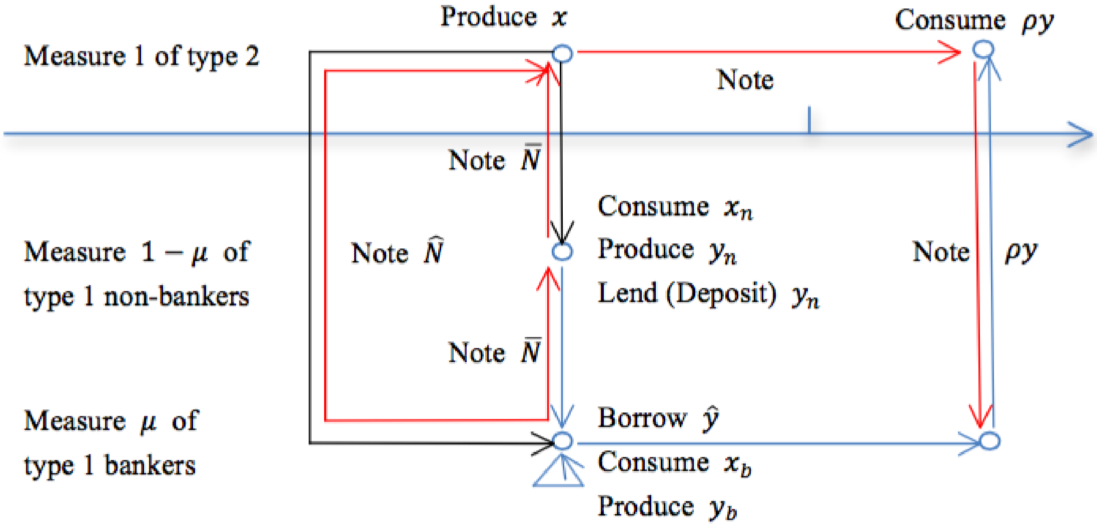

Figure 5: Timeline of decentralization with banking

Let V^1_{bt} be the banker's value function at time t given an allocation (x_{bt}, y_{bt}), which specifies that the banker consumes x_{bt} and produces y_{bt}, then the Bellman equation for the banker is

$$V^1_{bt} = U^1(x_{bt}, y_{bt}) + \beta V^1_{bt+1}. \tag{11}$$

Similarly, the bellman equations for type 1 non-banker and type 2 agent are, respectively,

$$V^1_{nt} = U^1(x_{nt}, y_{nt}) + \beta V^1_{nt+1}, \tag{12}$$

$$V^2_t = U^2(\rho y_t, x_t) + \beta V^2_{t+1}. \tag{13}$$

The repayment constraint for the banker is

$$\lambda\rho\left(\hat{y}_t + y_{bt}\right) + (1 - \Pi)\,\beta V^1_{bt+1} \leq \beta V^1_{bt+1}. \tag{14}$$

The LHS is the payoff from following the recommendation while the RHS is the deviation payoff. It reduces to

$$\hat{y}_t + y_{bt} \leq \frac{\beta\Pi}{\lambda\rho} V^1_{bt+1}. \tag{15}$$

By difining $\phi_t \equiv \frac{\beta\Pi}{\lambda\rho} V^1_{bt+1}$ as the debt limit, it is convenient to rewrite the repayment constraint as

$$\hat{y}_t + y_{bt} \leq \phi_t. \tag{16}$$

Using the bellman equation (11), we can express this recursively to make it clear that the debt limit in one period depends on the debt limit in the next period:

$$\phi_{t-1} = \frac{\beta\Pi}{\lambda\rho} U^1(x_{bt}, y_{bt}) + \beta\phi_t \tag{17}$$

There are a large number of spatially distinct Walrasian markets, and the agents trade short-term (across subperiod) credit contracts taking prices as given. Let goods y in the second subperiod be numeraire, the price of goods x in the first subperiod is p_{xt}, and the price of goods y in the first subperiod is p_{yt}. The banker maximizes utility given his budget constraint and repayment constraint. We drop the participation constraint because autarky is always feasible, and use the same preference functions as in the efficiency part.

$$\begin{aligned} &\max_{(x_{bt},y_{bt},\hat{y}_t)} u(x_{bt}) - y_{bt} - \tau \\ \text{s.t.} \quad & p_{xt}x_{bt} + p_{yt}\hat{y}_t = \rho(\hat{y}_t + y_{b_t}) \\ & \hat{y}_t + y_{bt} \leq \phi_t \end{aligned} \tag{18}$$

where $\rho > 1$.

Type 1 non-banker maximizes utility given his budget constraint.

$$\max_{(x_{nt},y_{nt})} u(x_{nt}) - y_{nt} + t \quad \text{s.t.} \quad p_{xt}x_{nt} = p_{yt}y_{nt} \tag{19}$$

Type 2 agent maximizes utility given his budget constraint.

$$\max_{(x_t,y_t)} \rho y_t - v(x_t) - \mu_t(k_0 + \Pi k) \quad \text{s.t.} \quad \rho y_t = p_{xt}x_t \tag{20}$$

Notice that the monitoring probability Π is exogenous in the decentralization, because otherwise, there will be a free-rider problem here. The cost $\mu_t(k_0 + \Pi k)$ is kind of a tax on type 2 to be used by the “government” to pay monitoring.

We have the following goods market clearing conditions: For goods y in the first subperiod, we have

$$\mu_t\hat{y}_t = (1 - \mu_t)\, y_{nt.} \tag{21}$$

For goods y in the second subperiod, we have

$$\rho y_t = \mu_t\rho\hat{y}_t + \mu_t\rho y_{bt}. \tag{22}$$

For goods x in the first subperiod, we have

$$\mu_t x_{bt} + (1-\mu_t)\, x_{nt} = x_t. \tag{23}$$

Combining the first two conditions, we have

$$y_t = \mu_t y_{bt} + (1-\mu_t)\, y_{nt.} \tag{24}$$

The free entry conditions are

$$\begin{aligned} \mu_t = 0 \quad & \text{if} \quad u(x_{bt}) - y_{bt} - \tau < u(x_{nt}) - y_{nt} + t \\ \mu_t \in (0,1) \quad & \text{if} \quad u(x_{bt}) - y_{bt} - \tau = u(x_{nt}) - y_{nt} + t \\ \mu_t = 1 \quad & \text{if} \quad u(x_{bt}) - y_{bt} - \tau > u(x_{nt}) - y_{nt} + t \end{aligned} \tag{25}$$

Following Alvarez and Jermann (2000), for all t, the equilibrium debt limit ϕ_t is defined as follows: the banker is indifferent between repaying ϕ_t and defaulting. In any feasible allocation, payoffs, and hence ϕ_t , must be bounded (so, as in many other models, we rule out explosive bubbles). We can also bound (x_{bt}, y_{bt}, $\hat{y}_t$, x_{nt}, y_{nt}, x_t, y_t) without loss in generality. Hence we have the following definition:

Definition 1. An equilibrium is a specification of nonnegative and bounded sequences of quantities $\{x^e_{bt}, y^e_{bt}, \hat{y}^e_t, x^e_{nt}, y^e_{nt}, x^e_t, y^e_t\}_{t=1}^{\infty}$, prices $\{p^e_{xt}, p^e_{yt}\}_{t=1}^{\infty}$, measure of bankers $\{\mu^e_t\}_{t=1}^{\infty}$ and credit limits $\{\phi^e_t\}_{t=1}^{\infty}$ such that for all t

1. $(x^e_{bt}, y^e_{bt}, \hat{y}^e_t)$ solves the banker's problem given ϕ^e_t.

2. (x^e_{nt}, y^e_{nt}) solves the type 1 non-banker's problem.

3. (x^e_t, y^e_t) solves the type 2 agent's problem.

4. Markets clear.

5. Free entry.

6. ϕ^e_t solves the difference equation (17) given (x^e_{bt}, y^e_{bt}).

Solve the type 1 non-banker's problem, we have

$$u'(x^e_{nt}) = p_{xt}/p_{yt} \Rightarrow x^e_{nt} = u'^{-1}(p_{xt}/p_{yt}) \tag{26}$$

The demand of goods x for type 1 non-banker x^e_{nt} is decreasing in p_{xt}.

Solve the type 2 agent's problem, we have

$$p_{xt} = v'(x^e_t) \Rightarrow x^e_t = v'^{-1}(p_{xt}) \tag{27}$$

The supply of goods x for type 2 agent x^e_t is increasing in p_{xt}.

Lemma 2. *There is an equilibrium only if $p_{yt} \leq \rho$.*

Proof: See the Appendix.

This lemma says that for the banker, the return of borrowing is always larger than or equal to the cost in equilibrium.

Lemma 3. *When $p_{yt} = \rho$, there is an equilibrium with no banks (trade).*

Proof: See the Appendix.

This lemma says when the return of borrowing is equal to the cost, there is an equilibrium with no banks (trade) if we charge a tax on bankers and give a transfer to non-bankers.

Lemma 4. *When $p_{yt} < \rho$, the repayment constraint must bind, $\hat{y}_t^e + y_{bt}^e = \phi_t$.*

Proof: If not, the banker could increase $\hat{y}_t$ to increase the objective function. □

From the budget constraint and the binding repayment constraint for type 1 banker, we have

$$y_{bt} = \frac{p_{xt} x_{bt} - (\rho - p_{yt})\, \phi_t}{p_{yt}} \tag{28}$$

Lemma 5. *When $p_{yt} < \rho$, $y_{bt}^e = 0$.*

Proof: Since the return of borrowing is larger than the cost, the banker would like to borrow as much as possible and produce nothing. □

From $y_{bt}^e = \frac{p_{xt} x_{bt}^e - (\rho - p_{yt}) \phi_t}{p_{yt}} = 0$, we have $x_{bt}^e = (\rho - p_{yt})\, \phi_t / p_{xt}$. The demand of goods x for the banker x_{bt}^e is decreasing in p_{xt}.

The bellman equation (17) can be rewritten as

$$\phi_{t-1} = f(\phi_t) \equiv \begin{cases} \frac{\beta\Pi}{\lambda\rho}\left[u((\rho - p_{yt})\,\phi_t/p_{xt}) - \tau\right] + \beta\phi_t & \text{if } 0 < \phi_t < y_b^{**} + \hat{y}^{**} \\ \frac{\beta\Pi}{\lambda\rho}\left[u((x_b^{**}) - y_b^{**} - \tau\right] + \beta\phi_t & \text{if } \phi_t \geq y_b^{**} + \hat{y}^{**} \\ 0 & \text{if } \phi_t = 0 \end{cases} \tag{29}$$

where x_b^{**}, y_b^{**} and $\hat{y}^{**}$ denote equilibrium sollutions ignoring the repayment constraint. The dynamical system describes the evolution of the debt limit in terms of itself. The three cases represent the evolution when the repayment constraint is binding, not binding, and when the debt limit is zero respectively.[11] This system is forward looking, naturally, in the sense that the debt limit in one period depends on the debt limit in the next period.

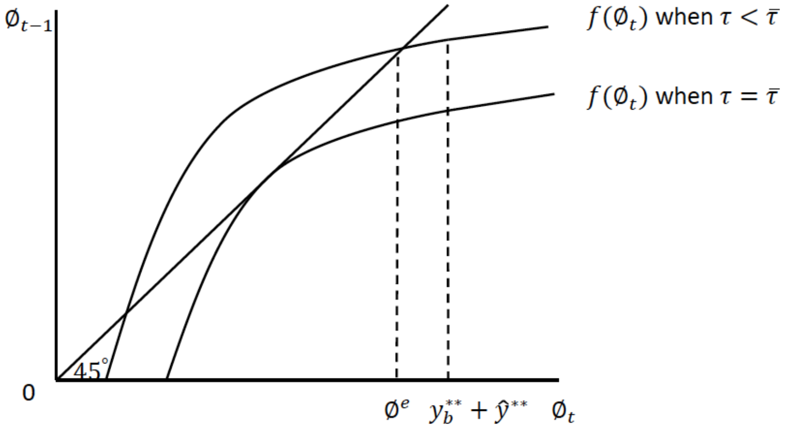

Figure 6: Steady state in terms of ϕ when $p_y < \rho$

[11] When the debt limit is zero, there is to be no credit in the future, you have nothing to lose by reneging, so no one will extend you credit today. Note that in this case there is no banker (trades), and no one needs to pay the tax.

A stationary equilibrium, or steady state, is a fixed point such that $f(\phi) = \phi$. Obviously $\phi = 0$ is one such point. A non-degenerate steady state is a solution to $f(\phi) = \phi > 0$. The graph of the steady state in terms of ϕ when $p_y < \rho$ is shown in Figure 6. The assumption that u satisfies Inada conditions $\lim\limits_{x \to 0} u'(x) = +\infty$ and $\lim\limits_{x \to +\infty} u'(x) = 0$ guaranties:

Proposition 5. *When $p_y < \rho$, if $0 < \tau < \bar{\tau}$, there are two stationary equilibriums with banks (trade), one is stable and the other is unstable; if $\tau = \bar{\tau}$, there is a unique stationary equilibrium with banks; if $\tau > \bar{\tau}$, there is an equilibrium with no banks.*

Proof: When $p_y < \rho$, the repayment constraint is binding. If $\tau = \bar{\tau}$, $\exists\ !\ \phi^e > 0$; if $\tau < \bar{\tau}$, $\exists$ two positive solutions. However, the one with larger ϕ^e such that $f'(\phi) < 1$ is stable, while the one with smaller ϕ^e such that $f'(\phi) > 1$ is unstable. The larger credit limit ϕ^e corresponds to a higher x_b^e and a higher payoff with $u'(x_b) < \frac{r\lambda\rho p_x}{\Pi(\rho - p_y)}$, and the smaller credit limit ϕ^e corresponds to a lower x_b^e and a lower payoff with $u'(x_b) > \frac{r\lambda\rho p_x}{\Pi(\rho - p_y)}$. □

When $\phi_{t-1} = \phi_b = \phi$, $x_{bt-1} = x_{bt} = x_b$, $p_{xt-1} = p_{xt} = p_x$ and $p_{yt-1} = p_{yt} = p_y$, The steady state condition regarding ϕ is $\phi = \frac{\beta\Pi}{\lambda\rho}\left[u((\rho - p_y)\,\phi/p_x) - \tau\right] + \beta\phi$, which reduces to

$$u(\frac{(\rho - p_y)\,\phi}{p_x}) = \frac{r\lambda\rho\phi}{\Pi} + \tau. \tag{30}$$

From $y_b = 0$, we have $\phi = x_b p_x / (\rho - p_y)$, thus the steady state condition regarding x_b is

$$u(x_b) = \frac{r\lambda\rho p_x x_b}{\Pi\,(\rho - p_y)} + \tau. \tag{31}$$

Use the market clearing conditions: For goods x

$$\mu x_b + (1-\mu)\, x_n = x \tag{32}$$

For goods y, $y = \mu y_b + (1-\mu)\, y_n$. Using the budget constraints, where $y = p_x x/\rho$ and $y_n = p_x x_n/p_y$, and $y_b = 0$, we have

$$x = (1-\mu)\, \rho x_n/p_y \tag{33}$$

Substituting (33) into (32), we have

$$\mu^e = \frac{(\rho/p_y - 1)\, x_n}{(\rho/p_y - 1)\, x_n + x_b} \in (0,1) \tag{34}$$

When $\tau \leq \bar{\tau}$, the equilibrium $(x_b^e, x_n^e, x^e, p_x^e, p_y^e, \mu^e)$ solves

$$\begin{aligned} u(x_b) &= \frac{r\lambda\rho p_x x_b}{\Pi\,(\rho - p_y)} + \tau \\ u'(x_n) &= p_x/p_y \\ p_x &= v'(x) \\ \mu x_b + (1-\mu)\, x_n &= x \\ x &= \ (1-\mu)\, \rho x_n/p_y \\ u(x_b) - \tau &= u(x_n) - p_x x_n + t \end{aligned}$$

where the first three equations are from the maximization problems of the type 1 banker, type 1 non-banker and type 2 agent; the fourth and fifth one are the market-clearing conditions for good x and good y, and the last one is the free-entry condition.

Compare it with the planner's problem, where $(x_b^*, x_n^*, x^*, \mu^*)$ solves

$$\begin{aligned}
u(x_b) &= \frac{r\lambda\rho k}{\rho - \frac{\theta}{1-\theta}} \\
u'(x_n) &= \frac{1-\theta}{\theta} v'(x) \\
\mu x_b + (1-\mu)\, x_n &= x \\
u(x_b) - u(x_n) + \frac{1-\theta}{\theta} \frac{u'(x_n)(x - x_b)}{1-\mu} &= \frac{1-\theta}{\theta} k_0
\end{aligned}$$

Proposition 6. *The competitive equilibrium is Pareto optimal if the following conditions are satisfied: (i)* $p_y^e = \frac{\theta}{1-\theta} < \rho$*, and (ii)* $\tau = \tau^*$*, where* τ^* *solves* $\frac{p_x^e(\tau) x_b^e(\tau)}{\Pi} + \frac{\tau(\rho - \frac{\theta}{1-\theta})}{r\lambda\rho} = k$*, and (iii)* $t = t^*$*, where* t^* *solves* $u\left[x_b^e(t)\right] - u\left[x_n^e(t)\right] + \frac{1-\theta}{\theta} \frac{u'\left[x_n^e(t)\right]\left[x^e(t) - x_b^e(t)\right]}{1-\mu^e(t)} = \frac{1-\theta}{\theta} k_0$.

Proof: See the Appendix.

For a given tax, we can have too much or too little entry, compared with the efficient outcome. If the tax is almost zero, nearly everyone wants to be a banker, and, thus, there is too much entry. If the tax is almost at the cut-off value, there is too little entry. When the government charges an optimal tax on the banker and gives an optimal transfer to the non-banker, the competitive equilibrium is efficient.

5.2 Rationing Bank Charters

The government can also impose a quota by limiting the number of bank charters at the efficiency level μ^*. In this way, we don't have the free entry condition and there is excess demand. A lottery is the easiest way to do the rationing scheme.

The equilibrium $(x_b^e, x_n^e, x^e, p_x^e, p_y^e)$ solves

$$\begin{aligned}
u(x_b) &= \frac{r\lambda\rho p_x x_b}{\Pi\left(\rho - p_y\right)} \\
u'(x_n) &= p_x/p_y \\
p_x &= v'(x) \\
\mu^* x_b + \left(1-\mu^*\right) x_n &= x \\
x &= \left(1-\mu^*\right)\rho x_n/p_y
\end{aligned}$$

Proposition 7. *The competitive equilibrium is Pareto optimal if the following conditions are satisfied: (i)* $p_y^e = \frac{\theta}{1-\theta} < \rho$*, and (ii)* $\mu = \mu^*$*, where* μ^* *is the efficient number of bankers.*

6 Conclusion

I develop a theoretical model with limited commitment and endogenous monitoring to study the optimal number and size of bankers from the planner's point of view. I begin by specifying preferences, technologies, and frictions, then illustrate how it can be desirable to designate some part of the ex-ante homogeneous agents to perform certain functions resembling banking: they accept deposits, they make investment, and their liabilities facilitate third party transactions. The mechanism is that if we have a utility cost to monitor the bankers, we can consider the cost-benefit trade-off of decreasing the number of bankers. Having fewer bankers reduces total monitoring cost, but for a given amount of total deposits, this means more deposits per banker. Having more deposits, however, increases the bankers' incentives to divert deposits for their own profit and thereby, reduces the benefit to the economy. The result is that the planner needs to give the bankers some reward to dissuade such opportunistic behavior. The optimal number of banks is negatively related to the fixed and marginal monitoring costs, negatively related to impatience, negatively related to the temptation to default, but positively related to the rate of return.

To implement efficient allocations, there is a tension between equilibrium with free entry and having positive bank profit for incentive reasons. In the competitive equilibrium, when the tax on banks is not too high, there exist non-degenerate stationary equilibriums. The allocation is optimal only if the government limits entry of banks. One natural way is to charge a tax on bankers and give a transfer to nonbankers; another way is to simply impose a quota by limiting the number of bank charters.

Chapter 2

One-child Policy, Life Cycle Earnings and the Household Saving Puzzle in China

7 Motivation

Rising saving rate, prominent among young households, is a typical feature in China. This observation presents a puzzle, however, because the standard representative agent model implies low saving when a household anticipates high income growth.

Many reasons can be contributable to the slightly high saving rate in China, mainly including:

One is the polarization in income distribution. Consumption ratio of necessity by high-income residents is far lower than the average level, while it's the opposite case for saving rate. If income is distributed more to the high-income, it's more likely to cause higher saving rate.

Sound social security system is yet to establish, which forces people to tighten their current consumption. As the reform of medical system and social security system is moving forward, the urban and rural residents will foot their own endowment insurance that is previously covered by employers, and the expected future expenses will increase, causing high saving and low consumption.

Reform of educational system places extremely heavy burden on people. China's financial educational funds among GDP are far behind the average world level. Both urban and rural residents find it difficult to afford higher education charge, which accounts for a substantially large part of both urban and rural household income, so they have to save money ahead of time.

Single option of alternative financial assets and limited way for investment by urban and rural residents, especially for those from rural areas, lead them to accumulate funds by way of centralized saving.

Due to China's family-planning policy, the increase of elderly population and small tendency of family members cause more accumulation of funds for caring the elderly.

We propose a resolution of the puzzle by analyzing the impact of the one-child policy in China and the resulting flattening of age-earning profiles on its household saving behavior. The mechanism is that with the implementation of the "one child policy"", the initial human capital of each young worker who enters into the job market increases, which results in a decrease of on the job training of each of them, and thus a flattening of age-earning profiles. The flattened earnings profiles encourage younger cohorts to save more for consumption smoothing, and therefore provides a factor for explaining high saving rates of the young.

Moreover, we use Barro and Becker's endogenous discount factor, which assumes that the time preference rate is a function of the number of children. According to Barro and Becker, when we decrease the number of children, the return on asset falls, and thus the parents would invest more on their children's human capital, which results in a flattening of age-earning profiles as well.

Finally, we introduce probability of survival in the old-age. In 1970, there is approximately a 72% chance of living into retirement, but by 2009, this has increased to 88.5%. The rise of probability of survival increases the saving rate as well.

8 Stylized Facts

China's one-child policy was progressively implemented in the 1970s, and strictly enforced in the urban areas by the 1980s. Figure 7 (from Choukhmane, Coeurdacier and Jin, 2014) shows the evolution of the fertility rate for urban households based on Census data: the fertility rate was a bit above three (per household) before 1970, started to decline during the period of 1972-1980 when the one child policy was progressively implemented, and reached a value very close to one after its strict implementation by 1982.

Figure 8 (from Song and Yang, 2010) panel A plots the Chinese urban household disposable income from 1982 to 2007 in 2007 Yuan. Data source is China Statistical Yearbook (CSY). Panel B plots the Chinese urban household saving rate. The solid and dotted lines stand for data from CSY and Urban Household Surveys (UHS), respectively. Saving rate is equal to (disposable income – consumption expenditure)/disposable income. It shows that the aggregate household saving rate decreased in the 1980s and started to rise since the early 1990s.

In Figure 9 (from Song and Yang, 2010) panel A, the dotted and solid lines refer to the cross-sectional age-saving profiles averaged over 1992-1993 and 2006-

2007 (weighted by the number of observations in each age cell), respectively. Some age cells contain very limited number of observations; thus, they use the three-age moving average to minimize the effect of measurement error. In the 1992-1993 period, the saving rates were relatively flat before age 45 and then increased towards the retirement age. While in the 2006-2007 period, the saving rates of young household increased a lot and the age-saving profile turned to a U-shape. The line in Panel B plots the increase of the age-specific saving rate from 1992-1993 to 2006-2007 (namely, the difference between the two profiles in Panel A). The U-shape pattern is more pronounced. The rise in the saving rate of the young generation sharply contrasts the typical hump-shaped profile in developed economies.

Figure 10 (from Song and Yang, 2010) shows the Cross-sectional Life-Cycle Earnings Profiles in 1992 and 2007. The dotted line in figure 10 presents the cross-sectional relative age-earnings profiles in 1992-1993. The solid line in the figure presents the cross-sectional relative age-earnings profile in 2006-2007. Workers of age 42 is used as the reference group to compute relative earnings. The flattening of the earnings profiles in 2006-2007 is evident: individuals at age of 50 earn essentially the same as those at age of 30. Although earnings remain to be increasing for age below 30, the slope of the profile has flattened out.

However, the cross-sectional earnings profile can not represent the individual's earnings profile. Both the cohort and year effects can make the cross-sectional earnings profile different, while the earnings difference between any age cells along a cross-sectional earnings profile comes from a combination of cohort and age effects. In order to see the earning differences from cohort to cohort, we are going to look at the cohort earnings profile. Since our CHNS dataset consists of repeated cross-sectional rather than panel data, we can investigate this issue only by constructing synthetic cohorts. The definition of cohort followed by Beaudry and Green (2000) is the year when individuals turn 25. And we use the CHNS (China Health and Nutrition Survey) data to run the following regression used by Beaudry and Green (2000) and Kambourov and Manovskii (2005).

$$\begin{aligned} logy(i,t) =& \alpha_0 + \alpha_1 z(i) + \alpha_2 z(i)^2 + \alpha_3 z(i) x(i,t) \\ & + \kappa_1 x(i,t) + \kappa_2 x(i,t)^2 + \kappa_3 x(i,t)^3 + \alpha_4 log\hat{Y}(t) + \epsilon(i,t) \end{aligned}$$

where the dependent variable, $logy(i,t)$, is the log earnings for cohort i at year t. And the independent variables include the cohort entry year, $z(i)$, the square of cohort entry year, and the interaction term of cohort entry year with the age $x(i,t)$ of cohort i in year t, plus the polynomial of age. The $log\hat{Y}(t)$ is the detrended aggregate earnings for year t, defined by $\hat{Y}(t) = Y(t)/(1+g)^t$, where g is the average growth rate of annual earnings in our sample. The estimated results are reported in Table 1.

	Cohort	Cohort Sq.	Cohort*Age	Age	Age Sq.	Age Cube
logy	$.1325^{***}$	$-.0002^{***}$	$-.0006^{***}$	$.2842^{***}$	$-.0033^{***}$	$.0000^{***}$

Table 1: Regression on Cohort-Specific Age-Earning Profiles

When α_2 is close to zero, we may simply interpret α_1 as the growth rate of the starting earnings. The positive and significant coefficient on the linear cohort term α_1 shows a higher growth rate of entry level earnings for younger generations. Another key coefficient of interest, α_3, on the cohort-age interaction term is negative and significant at 1% level. That is to say, the younger generations are facing a flatter earnings profile as they age. So the later cohorts start with a higher earning growth rate, but their earnings will eventually grow at a lower rate.

Figure 11 shows that life expectancy has increased fairly substantially from 64.0 in 1974 to 73.1 in 2009. With a higher life expectancy, I expect that the young generation's pre-cautionary saving would be higher.

9 Data

I have the CHNS (China Health and Nutrition Survey) data, which covers nine provinces in China that vary substantially in geography, economic development, public resources, and health indicators. There are about 4,400 households in the overall survey, covering about 19,000 individuals. Follow-up levels are high, but families that migrate from one community to a new one are not followed. I have the detailed income, wage and education level data collected in 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009 and 2011.

10 Literature Review

A growing body of empirical research found flatter age-earning profiles for younger cohort in different countries. Beaudry and Green (2000) found flatter age-earning profiles of Canadian men for recent cohorts in comparison with older cohorts. Similarly, Keane and Prasad (2006) use Poland data from 1985 to 1996, when Poland experienced a dramatic change in its political and economic structures. During this transition period, they found that the return to human capital increases but the return to experience declines. That is to say, the younger cohorts with higher human capital earn more but their earning growth rate would be lower compared with the older cohort. Kambourov and Manovskii (2009) claims that they are the first one to document a significant flattening of life-cycle earnings profiles for the successive cohorts of male workers in the U.S. entering the labor market since the late 1960s.

This paper is related to Song and Yang (2010). Their paper found a flattened age-earnings profile in China. And facing this exogenous flattened age-earnings profile, young workers have the incentive to save more today to compensate for reduced earnings growth over their lifetimes. The key difference in our paper is that we have endogenous flattened age-earnings profile. We explain the flattened age-earnings profile through the following mechanism: the implementation of the one-child policy increases the education expenditures on children, which results in a higher human capital when those children join the job market. The higher human capital will lead to a decrease of on the job training, and thus flattening the age-earnings profile.

Choukhmane, Coeurdacier and Jin (2014) uses one-child policy to explain the Chinese high saving rate through the "transfer channel" and the "expenditure channel". In the "transfer channel", parents with fertility constraint will have less children, and thus less transfer from their children, so they will save more for their old age. In the "expenditure channel", parents save more because they have fewer children to bear. But in their paper they assume young generations will borrow up to a constant fraction of their future wages. In our paper, we are trying to explain the high saving rate of the young cohorts.

11 Model

11.1 Set-up

Consider an overlapping generations economy in which agents live for four periods, characterized by: childhood (k), young-age (y), middle-age (m), and old-age (o). An individual in period $t-1$ does not make decisions on his consumption in childhood. In period t, they go to work, and need to decide the time $n_{y,t}$ to spend on the job training. At the end of period t, the young agent then makes the decision on the number of children f_t to bear. In middle-age, in $t+1$, the agent chooses the amount of education expenditure $x_{k,t+1}$ used in the production of human capital of each of his children. In old-age, the agent consumes all available resources, which is financed by gross return on accumulated assets, $Ra_{m,t+1}$.

Preferences: An individual maximizes the life-time utility which includes the consumption $c_{i,t}$ at each age i and the benefits from having f_t children:

$$\begin{aligned} U_t =& log(c_{y,t}) + \beta log(c_{m,t+1}) + p\beta^2 log(c_{o,t+2}) \\ &+ g(f_t)\beta \left[\alpha_1 log(c_{k,t+1}) + \alpha_2 log(x_{k,t+1})\right] \end{aligned}$$

where p is the probability of survival, and $g(f_t) = f_t^{\epsilon}, 0 < \epsilon < 1$. $g(f_t)$ measures the degree of altruism.

Budget constraints: The sequence of budget constraints for an individual born in $t-1$ follows:

$$\begin{aligned} c_{y,t} + a_{y,t} &= w_{y,t}(1 - n_{y,t}) \\ c_{m,t+1} + f_t c_{k,t+1} + f_t x_{k,t+1} + a_{m,t+1} &= w_{m,t+1} + R a_{y,t} \\ c_{o,t+2} &= R a_{m,t+1} \end{aligned}$$

Individuals lend (or borrow) through bank deposits, earning a constant and exogenously given gross interest R.

Wage rates:

$$\begin{aligned} w_{y,t} &= e z_t h_{y,t}^{\alpha} \\ w_{m,t+1} &= z_{t+1} h_{m,t+1}^{\alpha} \end{aligned}$$

with experience $e < 1$ and productivity z_t.

Human capital formation:

$$h_{k,t-1} = h_{m,t-1}^{\gamma_1} \tag{35}$$

$$h_{y,t} = (1-\delta_h) h_{k,t-1} + (h_{k,t-1})^{\gamma_2} x_{k,t-1}^{\gamma_3} h_{m,t-1}^{\gamma_4} \tag{36}$$

$$h_{m,t+1} = (1-\delta_h) h_{y,t} + (n_{y,t} h_{y,t})^{\gamma_5} \tag{37}$$

We follow the Ben-Porath (1967) formulation of the human capital production technology. A kid is born with innate ability $h^{\gamma_1}_{m,t-1}$, which depends on the parent's ability. In equation (36), the young-age human capital is subject to depreciation δ_h, and depends on the human capital production when he was a kid taking education investment $x_{k,t-1}$ from his parents. Equation (37) describes that the middle-age human capital formation depends on the depreciation δ_h and the time $n_{y,t}$ he puts on the job training to produce human capital.

11.2 Household Decisions

Optimal Consumption Decisions:

$$\begin{aligned} c_{y,t} &= [1+\beta+(\alpha_1+\alpha_2)f_t^{\epsilon}\beta+p\beta^2]^{-1}\left[w_{y,t}(1-n_{y,t})+\frac{w_{m,t+1}}{R}\right] \qquad (38) \\ c_{m,t+1} &= R\beta c_{y,t} \\ c_{k,t+1} &= \alpha_1 R\beta f_t^{\epsilon-1} c_{y,t} \\ c_{o,t+2} &= pR^2\beta^2 c_{y,t} \end{aligned}$$

The assumption of log-utility implies that the optimal consumption is a constant fraction of the present value of lifetime resources, which include the wage earnings from young and middle age.

Lemma 6. $\frac{\partial c_{y,t}}{\partial f_t} < 0$, *which means having fewer children will result in an increase in young-age consumption.*

The one-child policy was strictly enforced at 1979. Immediately after the implementation of this policy, fertility rate decreases a lot, which results in an increase in the consumption and a decrease in saving. It is consistent with the decreasing trend of saving rate in the 1980s in Figure 1. We will focus on explaining the clearly rising trend of the saving rate since the early 1990s in the following paragraphs.

Education Investment Choice:

$$x_{k,t+1} = \alpha_2 R \beta f_t^{\epsilon-1} c_{y,t} \tag{39}$$

Lemma 7.

$$\frac{\partial x_{k,t+1}}{\partial f_t} < 0$$

$$\frac{\partial h_{y,t+2}}{\partial x_{k,t+1}} > 0, \; \frac{\partial h_{y,t+2}}{\partial f_t} < 0$$

The lemma says that having fewer children will increase the education expenditures on each child. In other words, if there is a constraint on fertility rates, like one-child policy in China, people will have fewer children, and thus choose to put more resources on each kid's education expenditures used in the production of human capital. Then the next generation will have more human capital when they join the job market.

Fertility:

$$\epsilon \beta f_t^{\epsilon-1} R c_{y,t} \left[\alpha_1 log(c_{k,t+1}) + \alpha_2 log(x_{k,t+1})\right] = c_{k,t+1} + x_{k,t+1}$$

Fertility choice depends on equating the marginal utility of bearing an additional child with the net marginal cost of raising the child.

Job Training Decision and Human Capital:

$$n_{y,t}^{1-\gamma_5} = \frac{\alpha\gamma_5}{eR}\frac{z_{t+1}}{z_t}h_{m,t+1}^{\alpha-1}h_{y,t}^{\gamma_5-\alpha} \tag{40}$$

The job training decision hinges on equating the marginal benefit of gaining more middle-age human capital by training more with the marginal cost of losing the young-age working time.

Lemma 8. $\frac{\partial n_{y,t}}{\partial h_{y,t}} < 0$, *which says that if the young man joins the job market with a higher human capital, he will choose less job trainings.*

Age-earning Profile:

$$\frac{w_{m,t+1}}{w_{y,t}} = \frac{z_{t+1}}{ez_t}\left[(1-\delta_h) + n_{y,t}^{\gamma_5}h_{y,t}^{\gamma_5-1}\right]^{\alpha} \tag{41}$$

Lemma 9. *From Lemma 8, we can get:*

$$\frac{\partial w_{m,t+1}/w_{y,t}}{\partial h_{y,t}} < 0$$

which implies that if the young generation joins the market with a higher human capital, the age-earning profile will be flatter.

Saving Rates:

$$s_{y,t} = 1 - \frac{1}{[1 + \beta + (\alpha_1 + \alpha_2)\beta f_t^{\epsilon} + p\beta^2]} \left[\frac{1}{R(1 - n_{y,t})} \frac{w_{m,t+1}}{w_{y,t}} + 1 \right] \tag{42}$$

$$s_{m,t+1} = \frac{pR\beta^2}{[1 + \beta + (\alpha_1 + \alpha_2)\beta f_t^{\epsilon} + p\beta^2]} \left[\frac{1}{R} + (1 - n_{y,t}) \left(\frac{w_{m,t+1}}{w_{y,t}} \right)^{-1} \right] \tag{43}$$

Lemma 10. *From Lemma 8 and Lemma 9, we know that*

$$\frac{\partial s_{y,t}}{\partial w_{m,t+1}/w_{y,t}} < 0, \quad \frac{\partial s_{y,t}}{\partial p} > 0$$
$$\frac{\partial s_{m,t+1}}{\partial w_{m,t+1}/w_{y,t}} < 0, \quad \frac{\partial s_{m,t+1}}{\partial p} > 0$$

Lemma 10 tells us that facing a flatter age-earning profile, the young cohort will save more. It also implys that with a higher survival rate, the life expectancy would be higher, then the young cohort's pre-cautionary saving would be larger too.

Proposition 8. *Lemma 6-10 prove that our mechanism is valid. With the implementation of the one-child policy, education expenditures on young generation would increase, then the initial human capital of each young worker who enters into the job market increases, which results in a decrease of on the job training of each of them, and thus flattening the age-earning profiles. The flattened earnings profiles encourage younger cohorts to save more for consumption smoothing.*

12 A Quantitative OLG Model

Set-up and Model Dynamics

In this paper, t stands for the time period and a stands for age. For any variable x, $x_{a,t}$ represents the relevant variable for an individual with age a at time t. The economy is populated by overlapping generations of individuals who can live up to T period. After birth at time t, an individual lives with his parents until he is I years old, then he is independent, creates his own family and works. At age B, he has f_t chilidren. His children are going to be independent at his age $B+I$.

The dynamic programming problem for an individual with age a at time t, who has $H_{a,t}$ units of human capital and $K_{a,t}$ units of physical capital at time t, is given by the choice of life-cycle consumption path $\{c_{a+i,t+i}\}_{i=I}^{T}$, investment in children's consumption and education $\{C^k_{i,t}, X^k_{i,t}\}_{i=6}^{I-1}$, fertility rate f_t to solve:

$$V_{a,t}(H_{a,t}, K_{a,t}, n_{a,t}) = max \begin{bmatrix} U(C_{a,t}) + g(f_t)U_k(C^k_{a-B,t}, X^k_{a-B,t})I(B \leq a \leq B+I) \\ +\beta V_{a+1,t+1}(H_{a+1,t+1}, K_{a+1,t+1}, n_{a+1,t+1}) \end{bmatrix}$$

where U is strictly concave and increasing. β is a time preference discount factor. This function is maximized subject to the budget constraint

$$\begin{aligned} C_{a,t} + K_{a+1,t+1} + I(B \leq a \leq I)f_t(C^k_{a-B,t} + X^k_{a-B,t}) \leq & w_t(H_{a,t})(1-n_{a,t}) \\ & + (1+r_t)K_{a,t} \end{aligned}$$

where w_t is the wage rate at time t. The evolution of human capital for the parent

$$H_{a+1,t+1} = z(n_{a,t}H_{a,t})^{\gamma_5} + (1-\delta_h)H_{a,t}, \quad a \in [I, ..., R]$$

the evolution of human capital for the child

$$H^k_{a+1,t+1} = z(H^k_{a,t})^{\gamma_2}(X^k_{a,t})^{\gamma_3}(H^p_{B+a,t})^{\gamma_4} + (1-\delta_h)H^k_{a,t}, \quad a \in [6, I]$$

a time constraint on the job training decisions:

$$0 \le n_{a,t} \le 1, \text{ for all } a \text{ and } t$$

the child's initial stock of human capital is given by:

$$H^k_{6,t} = (H^p_{B,t})^{\gamma_1}$$

FOCs:

$$U^{'}(C_{a,t}) = \beta(1 + r_{t+1})U^{'}(C_{a+1,t+1})$$

$$(C^k_{a-B,t} + X^k_{a-B,t})U^{'}(C_{a,t}) = g^{'}(f_t)U_k(C^k_{a-B,t}, X^k_{a-B,t}), \quad \text{for } B \le a \le B + I$$

$$U^{'}(C_{a,t})f_t = g(f_t)U_{kc}(C^k_{a-B,t}, X^k_{a-B,t})$$

$$U^{'}(C_{a,t})f_t = g(f_t)U_{kx}(C^k_{a-B,t}, X^k_{a-B,t})$$

$$(n_{a,t}h_{a,t})^{1-\gamma_5} = \frac{w_{t+1}}{w_t(1+r_t)}[z\gamma_5 + (1-\delta_h)(n_{a+1,t+1}H_{a+1,t+1})^{1-\gamma_5}]$$

13 Calibration

To calibrate this model, We need the initial distribution across agents about their human capital stock as well as their parents' human capital level, age and capital stock. However, such information is not available from the data. Since the terminal on the job traning at retrie age would be 0, our strategy is to solve the human capital model backward. Rather than parameterizing the initial human capital, we parameterize the human capital at retire age H_r. Because there is a one to one relationship between initial human capital and human capital at retire age.

Then for any set of parameters $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \delta_h, H_{i,r})$, we can simulate the model and get the wage profiles as a function of those parameters. Then using nonlinear least squares to minimize over individuals:

$$\sum_i \sum_a (W^*_{i,a} - W_{i,a}(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \delta_h, H_r))^2$$

After getting the initial distribution across agent, then we can continue to solve the model starting from 1979 up to now ,and see whether the saving rates, schooling choices and age-earning profiles match the data or not.

14 Result

Assuming their is no technology change and the wage rate w_t is constant. So far, we have $\gamma_5 = 0.82$ and $\delta_h = 0.00066$ from the nonlinear least square calibration.

Figure 12 shows the simulated age-earning profiles for young cohort and old cohort. The blue solid line shows the average age-earning profiles for the cohort who were born in 1951-1956. The red solid line represents the average age-earning profiles for the cohort who were born in 1991-1996. I demeaned the red line to make the mean of the red line equal to the mean of the blue line. It is easy to see the flattening age-earning profile for the later cohort.

15 Conclusion

Using an overlapping generations model, we propose a resolution of the high household saving puzzle in China by analyzing the impact of the one-child policy and the resulting flattening of age-earning profiles on household saving behavior. Following Ben-Porath's (1967) human capital accumulation technology, with the implementation of the one-child policy, the initial human capital of each young worker who enters into the job market increases, which results in a decrease of the worker's on-the-job-training, and thus a flattening of age-earning profiles. The flattened age-earning profiles encourage younger cohorts to save more for consumption smoothing, and, therefore, provides an explanation for the high saving rates among the young. Moreover, we use the endogenous discount factor of Barro and Becker (1989), which assumes the time preference rate is a function of the number of children, and this amplifies the flattening of age-earning profiles. Both the data and the model demonstrate that our mechanism is valid.

References

[1] Acemoglu, D., Johnson, S. and Robinson, J. A., (2001). “The Colonial Origins of Comparative Development: An Empirical Investigation”. *American Economic Review* 91(5): 1369-1401. Alvarez, F., and Jermann, U. (2000), “Efficiency, Equilibrium, and Asset Pricing with Risk of Default”, *Econometrica*, 68, 775–798.

[2] Alvarez, F., and Jermann, U. (2000), “Efficiency, Equilibrium, and Asset Pricing with Risk of Default”, *Econometrica*, 68, 775–798.

[3] Andolfatto, D. and Nosal, E. (2009), “Money, Intermediation and Banking”, *Journal of Monetary Economics*, 56, 289–294.

[4] Arrow, Kenneth J. (1962). “The Economic Implications of Learning by Doing”. *The Review of Economic Studies*, Vol. 29, Issue 3, 155-173.

[5] Barro, R. J. and Becker, G. S. (1989) “Fertility Choice in a Model of Economic Growth”, *Econometrica*, S7(2), 481-501.

[6] Beaudry, P. and David A. G. (2000) “Cohort Patterns in Canadian earnings: Assessing the Role of Skill Premia in Inequality Trends”, *Canadian Journal of Economics*, 33(4), 907-936.

[7] Becker, G. S., Murphy, K. M. and Tamura, R. (1990). “Human Capital, Fertility, and Economic Growth”. *Journal of Political Economy*, University of Chicago Press, vol. 98(5), S12-S37.

[8] Becker, G. S. and Tomes, Nigel (1986) “Human Capital and the Rise and Fall of Families”. *Journal of Labor Economics* 4: S1 -S39.

[9] Ben-Porath, Y. (1967) “The Production of Human Capital and the Life Cycle of Earnings”, *Journal of Political Economy*, 75, 352-365.

[10] Berger, A. N., Kashyap, A. K. and Scalise, J. M. (1995), “The Transformation of the U.S. Banking Industry: What a Long, Strange Trip it’s been”, *Brookings Papers on Economic Activity,* 2, 55–201.

[11] Biais, B., Mariotti, T., Plantin, G. and Rochet, J-C. ((2007), “Dynamic Security Design: Convergence to Continuous Time and Asset Pricing Implications”, *Review of Economic Studies*, 74, 345–390.

[12] Boyd, J. and Prescott, E. (1986), “Financial Intermediary Coalitions”, *Journal of Economic Theory*, 38, 211–232.

[13] Broecker, T. (1990), “Credit Worthiness Tests and Interbank Competition”, *Econometrica*, 58, 429-452.

[14] Cavalcanti, R. and Wallace, N. (1999a), “A Model of Private Bank Note Issue”, *Review of Economic Dynamics*, 2, 104–136.

[15] Cavalcanti, R. and Wallace, N. (1999b), “Inside and Outside Money as Alternative Media of Exchange”, *Journal of Money, Credit, and Banking*, 31, 443–457.

[16] Choukhmane T., Coeurdacier N. and Jin, K. (2014) “The one-child policy and household savings”, *mimeo.*

[17] Corbae, D. and D'Erasmo, P. (2013) "A Quantitative Model of Banking Industry Dynamics", *mimeo.*

[18] Davies, G. (2002), *A History of Money From Ancient Times to the Present Day*, 3rd edn (Cardiff: University of Wales Press).

[19] Demarzo, P. and Fishman, M. (2007), "Agency and Optimal Investment Dynamics", *Review of Financial Studies*, 20, 151–188.

[20] Diamond, D. and Dybvig, P. (1983), "Bank Runs, Deposit Insurance, and Liquidity", *Journal of Political Economy*, 91, 401–419.

[21] Diamond, D. (1984), "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies,* 51, 393–414.

[22] Diamond, D. and Rajan, R. (2001), "Liquidity Risk, Liquidity Creation and Financial Fragility: A Theory of Banking," *Journal of Political Economy*, 109, 287–327.

[23] Ferguson, N. (2008), *The Ascent of Money* (New York: The Penguin Press).

[24] Freixas, X. and Rochet, J. C. (2008), *Microeconomics of Banking* (Cambridge: MIT Press).

[25] Gorton, G. and Winton, A. (2002), "Financial Intermediation", in Constantinides, G., Harris, M. and Stulz, R. (eds) *Handbook of the Economics of Finance* (Amsterdam/Boston: Elsevier/North Holland).

[26] Gu, C., Mattesini, F., Monnet, C. and Wright, R. (2013a), "Endogenous Credit Cycles", *Journal of Political Economy*, 121, 940–965.

[27] Gu, C., Mattesini, F., Monnet, C. and Wright, R. (2013b), "Banking: A New Monetarist Approach", *Review of Economic Studies,* 80(2): 636–662.

[28] He, P., Huang, L. and Wright, R. (2005), "Money and Banking in Search Equilibrium", *International Economic Review,* 46, 637–670.

[29] He, P., Huang, L. and Wright, R. (2008), "Money, Banking and Monetary Policy", *Journal of Monetary Economics,* 55, 1013–1024.

[30] Heckman, J., Lochner, L. and Taber, C. (1998). "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents". *Review of Economic Dynamics* 1, 1-58.

[31] Holmstrom, B., and Tirole, J. (1997), "Financial Intermediation, Loanable Funds, and the Real Sector", *Quarterly Journal of Economics,* 112(3), 663–691.

[32] Huangfu, X. and Sun, H. (2011), "Private Money and Bank Runs", *Canadian Journal of Economics*, 44, 859–879.

[33] Janicki, H. and Prescott, E. (2006) "Changes in the Size Distribution of U.S. Banks: 1960-2005", *Federal Reserve Bank of Richmond Quarterly Review*, 92, 291–316.

[34] Kambourov, G. and Manovskii, I. (2009) "Occupational Specificity of Human Capital", *International Economic Review* 50, 63-115.

[35] Keane, M. P., and Prasad, E. S. (2006) “Changes in the Structure of Earnings During the Polish Transition”, *Journal of Development Economics*, 80(2): 389-427.

[36] Kehoe, T. J., and Levine, D. K. (1993), “Debt-Constrained Asset Markets”, *Review of Economic Studies,* 60, 865–888.

[37] Kocherlakota, N. R. (1998), “Money Is Memory”, *Journal of Economic Theory*, 81, 232–251.

[38] Koeppl, T., Monnet, C. and Temzelides, T. (2008), “A Dynamic Model of Settlement”, *Journal of Economic Theory*, 142, 233–246.

[39] Lee, S. Y., Roys, N. and Seshadri, A. (2015) “The Causal Effect of Parental Human Capital on Children’s Human Capital”. mimeo, University of Wisconsin-Madison.

[40] Leland, H. E. and Pyle, D. H. (1977), “Informational Asymmetries, Financial Structure and Financial Intermediation”, *Journal of Finance*, 32, 371–387.

[41] Manuelli, R. E. and Seshadri, A. (2009). “Explaining International Fertility Differences”. *The Quarterly Journal of Economics*, MIT Press, vol. 124(2), pages 771-807.

[42] Mills, D. (2008), “Imperfect Monitoring and the Discounting of Inside Money”, International Economic Review, 49, 737–754.

[43] Selgin, G. (2007), “Banking”, written for *Encyclopedia Britannica.*

[44] Song, Z. M., and Yang, D. T. (2010) “Life Cycle Earnings and Saving in a Fast-Growing Economy”, *mimeo.*

[45] Townsend, R. (1979), “Optimal contracts and competitive markets with costly state verification”, *Journal of Economic Theory*, 22, 265–293.

[46] Townsend, R. (1988), “Models as Economies”, Economic Journal, 98, 1–24.

[47] Wallace, N. (1998), “A Dictum for Monetary Theory”, *Federal Reserve Bank of Minneapolis Quarterly Review*, Winter, 20–26.

[48] Wallace, N. (2005), “From Private Banking to Central Banking: Ingredients of a Welfare Analysis”, *International Economic Review*, 46, 619–636.

[49] Williamson, S. (1986), “Costly Monitoring, Financial Intermediation and Equilibrium Credit Rationing”, *Journal of Monetary Economics*, 18, 159–179.

[50] Williamson, S. (1987), “Financial Intermediation, Business Failures, and Real Business Cycles”, *Journal of Political Economy*, 95, 1196–1216.

Appendix

Proof of Proposition 1:

From the first and third FOCs, we have $u'(x_b) = u'(x_n) - \frac{1-\theta}{\theta}\frac{r\lambda\rho yku'(x_b)}{\mu[u(x_b)-y_b]^2}$, thus,

$u'(x_b^*) < u'(x_n^*)$. Because $u''(.) < 0$, we have $x_b^* > x_n^*$. From the second FOC, we have $y_b^* = 0$. □

Proof of Proposition 2:

The maximization problem for choosing y is $\max\limits_{y} -\theta y + (1-\theta)\left[\rho y - \frac{r\lambda\rho k y}{u(x_b)}\right]$

$y =$ anything if $-1+\frac{1-\theta}{\theta}\left[\rho - \frac{r\lambda\rho k}{u(x_b)}\right] = 0 \Rightarrow u(x_b) = \frac{r\lambda k}{1-\frac{\theta}{1-\theta}\frac{1}{\rho}}$. Thus, $\frac{\partial x_b^*}{\partial k} = \frac{\partial \overline{x_b}}{\partial k} > 0$, $\frac{\partial x_b^*}{\partial r} = \frac{\partial \overline{x_b}}{\partial r} > 0$, $\frac{\partial x_b^*}{\partial \lambda} = \frac{\partial \overline{x_b}}{\partial \lambda} > 0$, and $\frac{\partial x_b^*}{\partial \rho} = \frac{\partial \overline{x_b}}{\partial \rho} < 0$. Then, from the other FOCs, we can solve y^*.

$y = +\infty$ if $-1 + \frac{1-\theta}{\theta}\left[\rho - \frac{r\lambda\rho k}{u(x_b)}\right] > 0$, it is not the solution.

$y = 0$ if $-1 + \frac{1-\theta}{\theta}\left[\rho - \frac{r\lambda\rho k}{u(x_b)}\right] < 0$, it is not the solution. □

Proof of Proposition 3:

Differentiation of the FOCs yields

$$\begin{bmatrix} -\frac{1-\theta}{\theta}v''(x)\,(1-\mu) & \frac{A}{y} & -\frac{1-\theta}{\theta}v''(x)(\overline{x_b}-x_n) - \frac{A}{\mu^2} \\ u''(x_n) - \frac{1-\theta}{\theta}v''(x)\,(1-\mu) & 0 & -\frac{1-\theta}{\theta}v''(x)(\overline{x_b}-x_n) \\ -\frac{1-\theta}{\theta}v''(x)(\overline{x_b}-x_n)\,(1-\mu) & 0 & -\frac{1-\theta}{\theta}v''(x)(\overline{x_b}-x_n)^2 \end{bmatrix} \begin{bmatrix} dx_n \\ dy \\ d\mu \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & \frac{A}{k} + B\frac{\partial\overline{x_b}}{\partial k} & \frac{A}{r} + B\frac{\partial\overline{x_b}}{\partial r} & \frac{A}{\lambda} + B\frac{\partial\overline{x_b}}{\partial \lambda} & \frac{A}{\rho} + B\frac{\partial\overline{x_b}}{\partial \rho} \\ 0 & C\frac{\partial\overline{x_b}}{\partial k} & C\frac{\partial\overline{x_b}}{\partial r} & C\frac{\partial\overline{x_b}}{\partial \lambda} & C\frac{\partial\overline{x_b}}{\partial \rho} \\ -1 & \Sigma\frac{\partial\overline{x_b}}{\partial k} & \Sigma\frac{\partial\overline{x_b}}{\partial r} & \Sigma\frac{\partial\overline{x_b}}{\partial \lambda} & \Sigma\frac{\partial\overline{x_b}}{\partial \rho} \end{bmatrix} \begin{bmatrix} dk_0 \\ dk \\ dr \\ d\lambda \\ d\rho \end{bmatrix}$$

=0,

where $A = \frac{1-\theta}{\theta}\frac{r\lambda\rho yku'(\overline{x_b})}{\mu[u(\overline{x_b})]^2}$, $B = u''(\overline{x_b}) - \frac{1-\theta}{\theta}v''(x)\mu + \frac{1-\theta}{\theta}\frac{r\lambda\rho yk}{\mu}\frac{u''(\overline{x_b})[u(\overline{x_b})]^2 - 2u(\overline{x_b})[u'(\overline{x_b})]^2}{[u(\overline{x_b})]^4}$, $C = -\frac{1-\theta}{\theta}v''(x)\mu$, $\Sigma = u'(\overline{x_b}) - u'(x_n) - \frac{1-\theta}{\theta}\mu v''(x)(\overline{x_b} - x_n)$. The determinant of the square matrix is

$$D = \frac{\left(\frac{1-\theta}{\theta}\right)^2 r\lambda\rho k u'(\overline{x_b})u''(x_n)v''(x)(\overline{x_b} - x_n)^2}{\mu\left[u(\overline{x_b})\right]^2} < 0.$$

The partial derivatives of x_n^* with respect to each of its arguments are, respectively,

$$\begin{aligned}
\frac{\partial x_n^*}{\partial k} &= \frac{\left(\frac{1-\theta}{\theta}\right)^2 \frac{r\lambda\rho k u'(\overline{x_b})}{\mu[u(\overline{x_b})]^2}\left\{\left[u'(\overline{x_b}) - u'(x_n)\right] v''(x)(\overline{x_b} - x_n)\right\}\frac{\partial \overline{x_b}}{\partial k}}{D} > 0,\\
\frac{\partial x_n^*}{\partial r} &= \frac{\left(\frac{1-\theta}{\theta}\right)^2 \frac{r\lambda\rho k u'(\overline{x_b})}{\mu[u(\overline{x_b})]^2}\left\{\left[u'(\overline{x_b}) - u'(x_n)\right] v''(x)(\overline{x_b} - x_n)\right\}\frac{\partial \overline{x_b}}{\partial r}}{D} > 0,\\
\frac{\partial x_n^*}{\partial \lambda} &= \frac{\left(\frac{1-\theta}{\theta}\right)^2 \frac{r\lambda\rho k u'(\overline{x_b})}{\mu[u(\overline{x_b})]^2}\left\{\left[u'(\overline{x_b}) - u'(x_n)\right] v''(x)(\overline{x_b} - x_n)\right\}\frac{\partial \overline{x_b}}{\partial \lambda}}{D} > 0,\\
\frac{\partial x_n^*}{\partial \rho} &= \frac{\left(\frac{1-\theta}{\theta}\right)^2 \frac{r\lambda\rho k u'(\overline{x_b})}{\mu[u(\overline{x_b})]^2}\left\{\left[u'(\overline{x_b}) - u'(x_n)\right] v''(x)(\overline{x_b} - x_n)\right\}\frac{\partial \overline{x_b}}{\partial \rho}}{D} < 0.
\end{aligned}$$

The partial derivatives of $x_b^* - x_n^*$ with respect to each of its arguments are, respectively,

$$\begin{aligned}
\frac{\partial\left(x_b^* - x_n^*\right)}{\partial k} &= \frac{\left(\frac{1-\theta}{\theta}\right)^2 \frac{r\lambda\rho k u'(\overline{x_b})}{\mu[u(\overline{x_b})]^2} v''(x)(\overline{x_b} - x_n)\Phi\frac{\partial \overline{x_b}}{\partial k}}{D} > 0,\\
\frac{\partial\left(x_b^* - x_n^*\right)}{\partial r} &= \frac{\left(\frac{1-\theta}{\theta}\right)^2 \frac{r\lambda\rho k u'(\overline{x_b})}{\mu[u(\overline{x_b})]^2} v''(x)(\overline{x_b} - x_n)\Phi\frac{\partial \overline{x_b}}{\partial r}}{D} > 0,\\
\frac{\partial\left(x_b^* - x_n^*\right)}{\partial \lambda} &= \frac{\left(\frac{1-\theta}{\theta}\right)^2 \frac{r\lambda\rho k u'(\overline{x_b})}{\mu[u(\overline{x_b})]^2} v''(x)(\overline{x_b} - x_n)\Phi\frac{\partial \overline{x_b}}{\partial \lambda}}{D} > 0,\\
\frac{\partial\left(x_b^* - x_n^*\right)}{\partial \rho} &= \frac{\left(\frac{1-\theta}{\theta}\right)^2 \frac{r\lambda\rho k u'(\overline{x_b})}{\mu[u(\overline{x_b})]^2} v''(x)(\overline{x_b} - x_n)\Phi\frac{\partial \overline{x_b}}{\partial \rho}}{D} < 0.
\end{aligned}$$

where $\Phi = u''(x_n)(\overline{x_b}-x_n)-[u'(\overline{x_b})-u'(x_n)]$. According to the mean value theorem, there exists a point ξ in $(x_n,\overline{x_b})$ such that $u''(\xi)=\frac{u'(\overline{x_b})-u'(x_n)}{\overline{x_b}-x_n}$, thus, for $u(x)=\frac{x^{1-\alpha}-1}{1-\alpha}$, where $\alpha>0$, we have $\Phi=[u''(x_n)-u''(\xi)](\overline{x_b}-x_n)<0$. □

Proof of Proposition 4:

The partial derivatives of μ^* with respect to each of its arguments are, respectively,

$$\frac{\partial\mu^*}{\partial k_0}=\frac{-\left(\frac{1-\theta}{\theta}\right)^2\frac{r\lambda\rho k u'(\overline{x_b})}{\mu[u(\overline{x_b})]^2}\left[u''(x_n)-\frac{1-\theta}{\theta}(1-\mu)v''(x)\right]}{D}<0,$$
$$\frac{\partial\mu^*}{\partial k}=\frac{-\frac{1-\theta}{\theta}\frac{r\lambda\rho k u'(\overline{x_b})\Omega}{\mu[u(\overline{x_b})]^2}\frac{\partial\overline{x_b}}{\partial k}}{D}<0,$$
$$\frac{\partial\mu^*}{\partial r}=\frac{-\frac{1-\theta}{\theta}\frac{r\lambda\rho k u'(\overline{x_b})\Omega}{\mu[u(\overline{x_b})]^2}\frac{\partial\overline{x_b}}{\partial r}}{D}<0,$$
$$\frac{\partial\mu^*}{\partial\lambda}=\frac{-\frac{1-\theta}{\theta}\frac{r\lambda\rho k u'(\overline{x_b})\Omega}{\mu[u(\overline{x_b})]^2}\frac{\partial\overline{x_b}}{\partial\lambda}}{D}<0,$$
$$\frac{\partial\mu^*}{\partial\rho}=\frac{-\frac{1-\theta}{\theta}\frac{r\lambda\rho k u'(\overline{x_b})\Omega}{\mu[u(\overline{x_b})]^2}\frac{\partial\overline{x_b}}{\partial\rho}}{D}>0,$$

where $\Omega=\frac{1-\theta}{\theta}(1-\mu)v''(x)[u'(\overline{x_b})-u'(x_n)]-u''(x_n)\left[u'(\overline{x_b})-u'(x_n)-\frac{1-\theta}{\theta}\mu v''(x)(\overline{x_b}-x_n)\right]$.

□

Proof of Lemma 2:

The Lagrangean function for the banker is:

$$\mathcal{L}=u(x_b)-y_b-\tau+\lambda_1[\rho(\hat{y}+y_b)-p_x x_b-p_y\hat{y}]+\lambda_2(\phi-\hat{y}-y_b)+\lambda_3 y_b.$$

The critical points of the Lagrangean are the solutions $(x_b,\, y_b,\, \hat{y},\, \lambda_1,\, \lambda_2,\, \lambda_3)$ to the following system of equations:

1. $u'(x_b) - \lambda_1 p_x = 0,$
2. $-1 + \lambda_1 \rho - \lambda_2 + \lambda_3 = 0,$
3. $\lambda_1 \rho - \lambda_1 p_y - \lambda_2 = 0$,
4. $\lambda_2 \geq 0,\, \phi - \hat{y} - y_b \geq 0,\, \lambda_2 \left(\phi - \hat{y} - y_b\right) = 0,$
5. $\lambda_3 \geq 0,\, y_b \geq 0,\, \lambda_3 y_b = 0,$

where the first three equations are the first order conditions for x_b, y_b, and $\hat{y}$, and the last two equations are the complementary slackness conditions. Because $\lambda_1 \left(\rho - p_y\right) = \lambda_2 \geq 0$, we have $p_y \leq \rho$. □

Proof of Lemma 3:

When $p_y = \rho$, we have $\lambda_2 = 0$.

If $y_b > 0$, we have $\lambda_3 = 0 \Rightarrow \lambda_1 = 1/\rho$, $u'(x_b) = p_x/\rho = p_x/p_y = u'(x_n) \Rightarrow x_b^e = x_n^e$, $y_b^e = y_n^e$. If we charge a tax on bankers and give a transfer to non-bankers, the bankers have a lower payoff than the non-bankers, there is no banks (trade).

If $y_b = 0$, from $p_y = \rho$, the budget constraint becomes $p_x x_b = \rho y_b$, thus $x_b = 0$, there is no banks (trade). □

Proof of Proposition 6:

Compare the results in the efficiency part and the equilibrium part, we have three different equations, the binding repayment constraint, the free-entry condition, and the optimal consumption relationship between the non-banker and type 2. We can prove the three conditions step by step:

(i) Prove $p_y^e = \frac{\theta}{1-\theta} < \rho$.

From the efficiency part, we have $u'(x_n) = \frac{1-\theta}{\theta} v'(x)$, while from the equilibrium part, we have $u'(x_n) = v'(x)/p_y$. We need to have $p_y = \frac{\theta}{1-\theta}$ such that the efficient allocations can be implemented.

(ii) Prove $\tau = \tau^*$, where τ^* solves $\frac{p_x^e(\tau)x_b^e(\tau)}{\Pi} + \frac{\tau(\rho - \frac{\theta}{1-\theta})}{r\lambda\rho} = k$.

The first equation in the efficiency part is

$$u(x_b) = \frac{r\lambda\rho k}{\rho - \frac{\theta}{1-\theta}},$$

while the first equation in the equilibrium can be rewritten as

$$u(x_b) = \frac{r\lambda\rho}{\rho - p_y}\left[\frac{p_x x_b}{\Pi} + \frac{\tau(\rho - p_y)}{r\lambda\rho}\right].$$

Using $p_y^e = \frac{\theta}{1-\theta}$, to let the banker's consumption in the equilibrium reach the optimal outcome, we need

$$\frac{p_x^e\left(\tau\right)x_b^e\left(\tau\right)}{\Pi} + \frac{\tau(\rho - \frac{\theta}{1-\theta})}{r\lambda\rho} = k.$$

(iii) Prove $t = t^*$, where t^* solves $u\left[x_b^e(t)\right] - u\left[x_n^e\left(t\right)\right] + \frac{1-\theta}{\theta}\frac{u'\left[x_n^e(t)\right]\left[x^e(t) - x_b^e(t)\right]}{1-\mu^e(t)} = \frac{1-\theta}{\theta}k_0$.

We need to set t to the optimal level such that the equilibrium allocations satisfy the last equation that is different in the efficiency part. □

Proof of Lemma 6:

From the equation (38)

$$c_{y,t} = [1 + \beta + (\alpha_1 + \alpha_2)f_t^{\epsilon}\beta + p\beta^2]^{-1}\left[w_{y,t}(1 - n_{y,t}) + \frac{w_{m,t+1}}{R}\right]$$

The wage profile will not be affected by the fertility choice f_t, the we have:

$$\frac{\partial c_{y,t}}{\partial f_t} = -\epsilon(\alpha_1 + \alpha_2)\beta f_t^{\epsilon-1}[1 + \beta + (\alpha_1 + \alpha_2)f_t^{\epsilon}\beta + p\beta^2]^{-2}\left[w_{y,t}(1 - n_{y,t}) + \frac{w_{m,t+1}}{R}\right] < 0$$

□

Proof of Lemma 7:

From the education investment choice (equation (39))

$$x_{k,t+1} = \alpha_2 R\beta f_t^{\epsilon-1}c_{y,t}$$

then we have:

$$\frac{\partial x_{k,t+1}}{\partial f_t} = (\epsilon - 1)\alpha_2 R\beta c_{y,t}f_t^{\epsilon-2} + \alpha_2 R\beta f_t^{\epsilon-1}\frac{\partial c_{y,t}}{\partial f_t} < 0$$

since $\epsilon < 1$ and $\frac{\partial c_{y,t}}{\partial f_t} < 0$ by Lemma 1.

By the human capital accumulation equation (36), it is easy to get:

$$\frac{\partial h_{y,t+2}}{\partial x_{k,t+1}} = \gamma_3 (h_{k,t+1})^{\gamma_2} x_{k,t+1}^{\gamma_3 - 1} h_{m,t+1}^{\gamma_4} > 0$$

So,

$$\frac{\partial h_{y,t+2}}{\partial f_t} = \frac{\partial h_{y,t+2}}{\partial x_{k,t+1}} \frac{\partial x_{k,t+1}}{\partial f_t} < 0$$

□

Proof of Lemma 8:

From equation (37) and (40)

$$\begin{aligned}
n_{y,t}^{1-\gamma_5} &= \frac{\alpha\gamma_5}{eR} \frac{z_{t+1}}{z_t} h_{m,t+1}^{\alpha-1} h_{y,t}^{\gamma_5-\alpha} \\
&= \frac{\alpha\gamma_5}{eR} \frac{z_{t+1}}{z_t} \left[(1-\delta_h) h_{y,t} + (n_{y,t} h_{y,t})^{\gamma_5}\right]^{\alpha-1} h_{y,t}^{\gamma_5-\alpha} \\
&= \frac{\alpha\gamma_5}{eR} \frac{z_{t+1}}{z_t} \left[(1-\delta_h) h_{y,t}^{\frac{\gamma_5-1}{\alpha-1}} + n_{y,t}^{\gamma_5} h_{y,t}^{\frac{\alpha(\gamma_5-1)}{\alpha-1}}\right]^{\alpha-1} \\
n_{y,t}^{\frac{1-\gamma_5}{\alpha-1}} &= \left(\frac{\alpha\gamma_5}{eR}\right)^{\frac{1}{\alpha-1}} \left(\frac{z_{t+1}}{z_t}\right)^{\frac{1}{\alpha-1}} \left[(1-\delta_h) h_{y,t}^{\frac{\gamma_5-1}{\alpha-1}} + n_{y,t}^{\gamma_5} h_{y,t}^{\frac{\alpha(\gamma_5-1)}{\alpha-1}}\right]
\end{aligned}$$

Then

$$\frac{\partial n_{y,t}}{\partial h_{y,t}} = \frac{\left(\frac{1-\gamma_5}{1-\alpha}\right) \left(\frac{\alpha\gamma_5}{eR}\right)^{\frac{1}{\alpha-1}} \left(\frac{z_{t+1}}{z_t}\right)^{\frac{1}{\alpha-1}} \left[(1-\delta_h) h_{y,t}^{\frac{\gamma_5-\alpha}{\alpha-1}} + \alpha n_{y,t}^{\gamma_5} h_{y,t}^{\frac{\alpha(\gamma_5-1)}{\alpha-1}-1}\right]}{\left(\frac{1-\gamma_5}{\alpha-1}\right) n_{y,t}^{\frac{2-\gamma_5-\alpha}{\alpha-1}} - \left(\frac{\alpha\gamma_5}{eR}\right)^{\frac{1}{\alpha-1}} \left(\frac{z_{t+1}}{z_t}\right)^{\frac{1}{\alpha-1}} \gamma_5 n_{y,t}^{\gamma_5-1} h_{y,t}^{\frac{\alpha(\gamma_5-1)}{\alpha-1}}}$$

Since $\gamma_5 < 1$ and $\alpha < 1$, then the numerator is larger than zero and the denominator is smaller than zero. So we have:

$$\frac{\partial n_{y,t}}{\partial h_{y,t}} < 0$$

□

Proof of Lemma 9:

From the age-earning profile equation (41), we can get:

$$\frac{\partial w_{m,t+1}/w_{y,t}}{\partial h_{y,t}} = \frac{\alpha z_{t+1}}{e z_t} \left[(1-\delta_h) + n_{y,t}^{\gamma_5} h_{y,t}^{\gamma_5 - 1}\right]^{\alpha-1} \left[\gamma_5 (n_{y,t} h_{y,t})^{\gamma_5 - 1} \frac{\partial n_{y,t}}{\partial h_{y,t}} + (\gamma_5 - 1) n_{y,t}^{\gamma_5} h_{y,t}^{\gamma_5 - 2}\right]$$

By Lemma 3 and $\gamma_5 < 1$, we have:

$$\frac{\partial w_{m,t+1}/w_{y,t}}{\partial h_{y,t}} < 0$$

□

Proof of Lemma 10:

By the young-age saving rate function (42):

$$s_{y,t} = 1 - \frac{1}{[1 + \beta + (\alpha_1 + \alpha_2)\beta f_t^{\epsilon} + p\beta^2]} \left[\frac{1}{R(1 - n_{y,t})} \frac{w_{m,t+1}}{w_{y,t}} + 1\right]$$

It is easy to get:

$$\begin{aligned}\frac{\partial s_{y,t}}{\partial \frac{w_{m,t+1}}{w_{y,t}}} &= -\frac{1}{R\left[1+\beta+(\alpha_1+\alpha_2)\beta f_t^{\epsilon}+p\beta^2\right](1-n_{y,t})} \\ &< 0\end{aligned}$$

By Lemma 4 $\frac{\partial w_{m,t+1}/w_{y,t}}{\partial h_{y,t}} < 0$, we have:

$$\frac{\partial s_{y,t}}{\partial h_{y,t}} = \frac{\partial s_{y,t}}{\partial w_{m,t+1}/w_{y,t}} \frac{\partial w_{m,t+1}/w_{y,t}}{\partial h_{y,t}} > 0$$

Similarly,

$$\frac{\partial s_{y,t}}{\partial p} = \frac{\beta^2}{\left[1+\beta+(\alpha_1+\alpha_2)\beta f_t^{\epsilon}+p\beta^2\right]^2}\left[\frac{1}{R(1-n_{y,t})}\frac{w_{m,t+1}}{w_{y,t}}+1\right] > 0$$

For the mid-age saving rate,

$$\frac{\partial s_{m,t+1}}{\partial w_{m,t+1}/w_{y,t}} = -\frac{pR\beta^2}{\left[1+\beta+(\alpha_1+\alpha_2)\beta f_t^{\epsilon}+p\beta^2\right]}(1-n_{y,t})\left(\frac{w_{m,t+1}}{w_{y,t}}\right)^{-2} < 0$$

So,

$$\frac{\partial s_{m,t+1}}{\partial h_{y,t}} = \frac{\partial s_{m,t+1}}{\partial w_{m,t+1}/w_{y,t}} \frac{\partial w_{m,t+1}/w_{y,t}}{\partial h_{y,t}} > 0$$

$$\frac{\partial s_{m,t+1}}{\partial p} = \frac{R\beta^2\left[1+\beta+(\alpha_1+\alpha_2)\beta f_t^{\epsilon}\right]}{\left[1+\beta+(\alpha_1+\alpha_2)\beta f_t^{\epsilon}+p\beta^2\right]^2}\left[\frac{1}{R}+(1-n_{y,t})\left(\frac{w_{m,t+1}}{w_{y,t}}\right)^{-1}\right] > 0$$

□

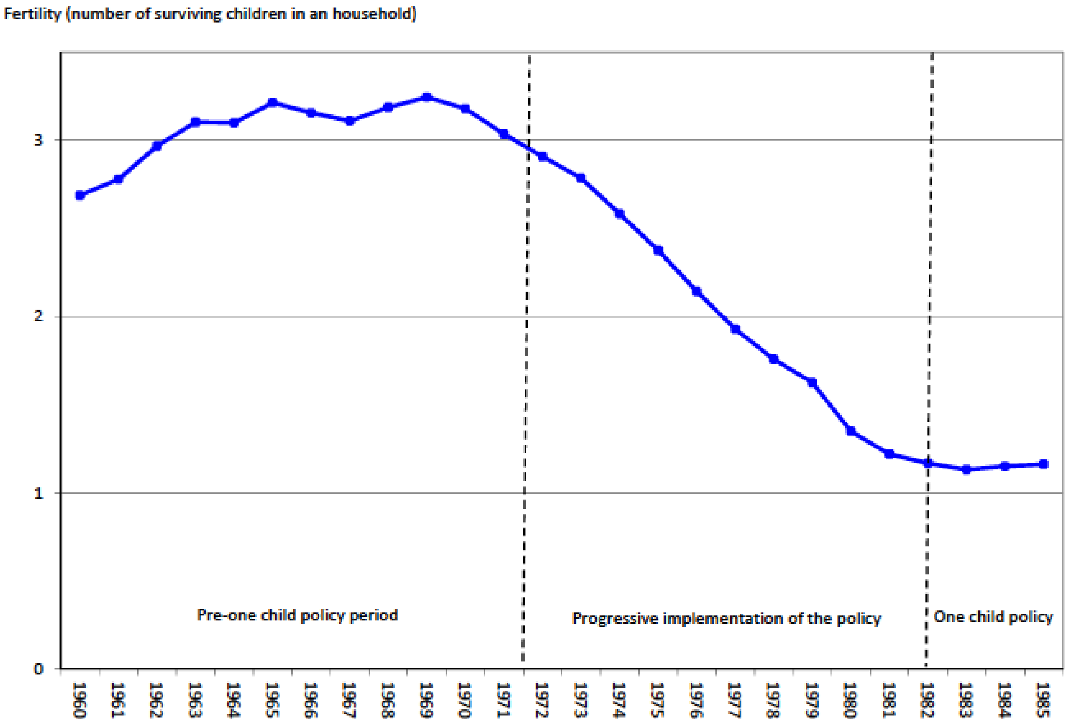

Figure 7: Fertility in Chinese Urban Areas

Notes: Data source: Census, restricted sample where only urban households are considered.

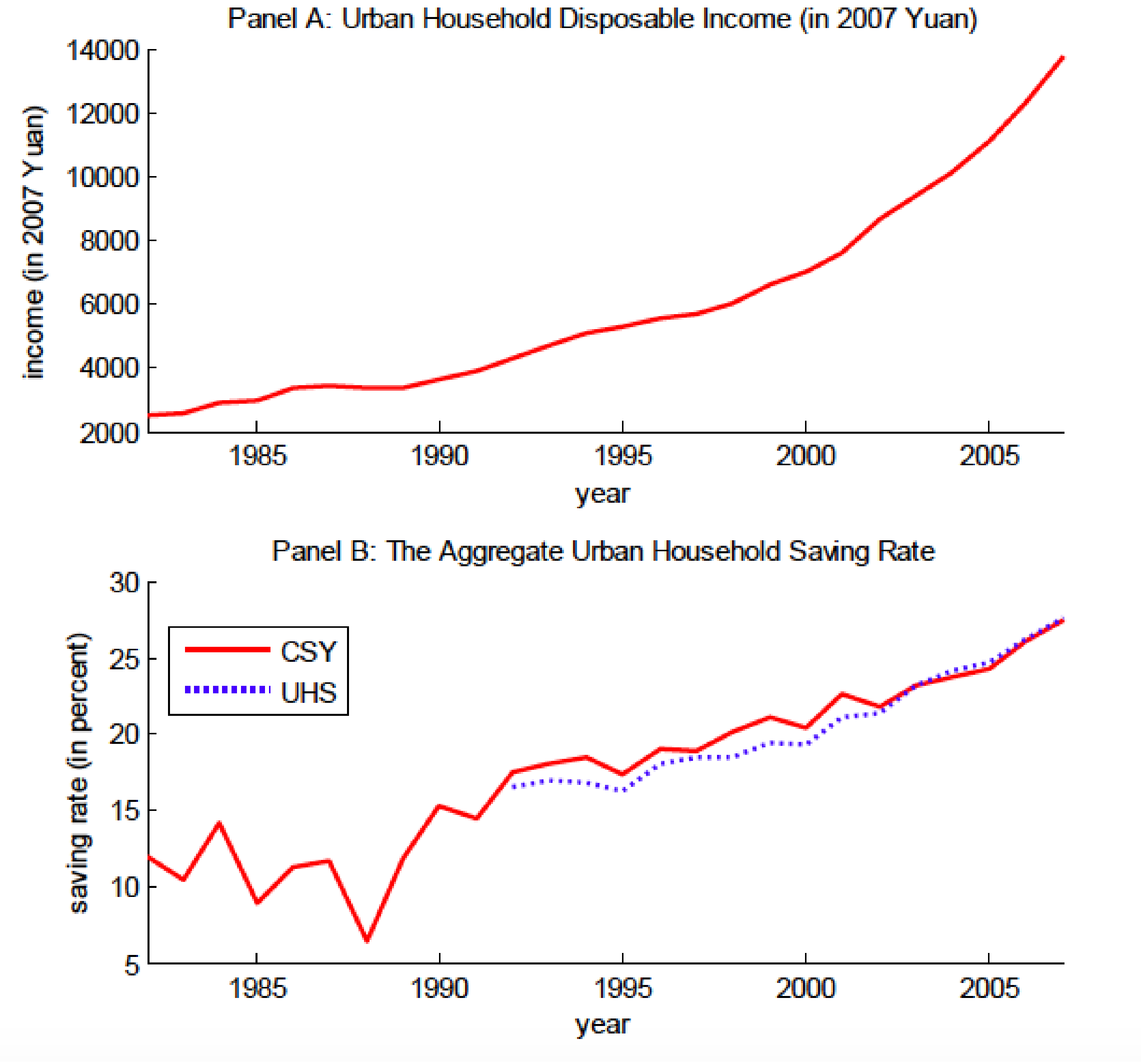

Figure 8: The Aggregate Urban Household Income and Saving Rate

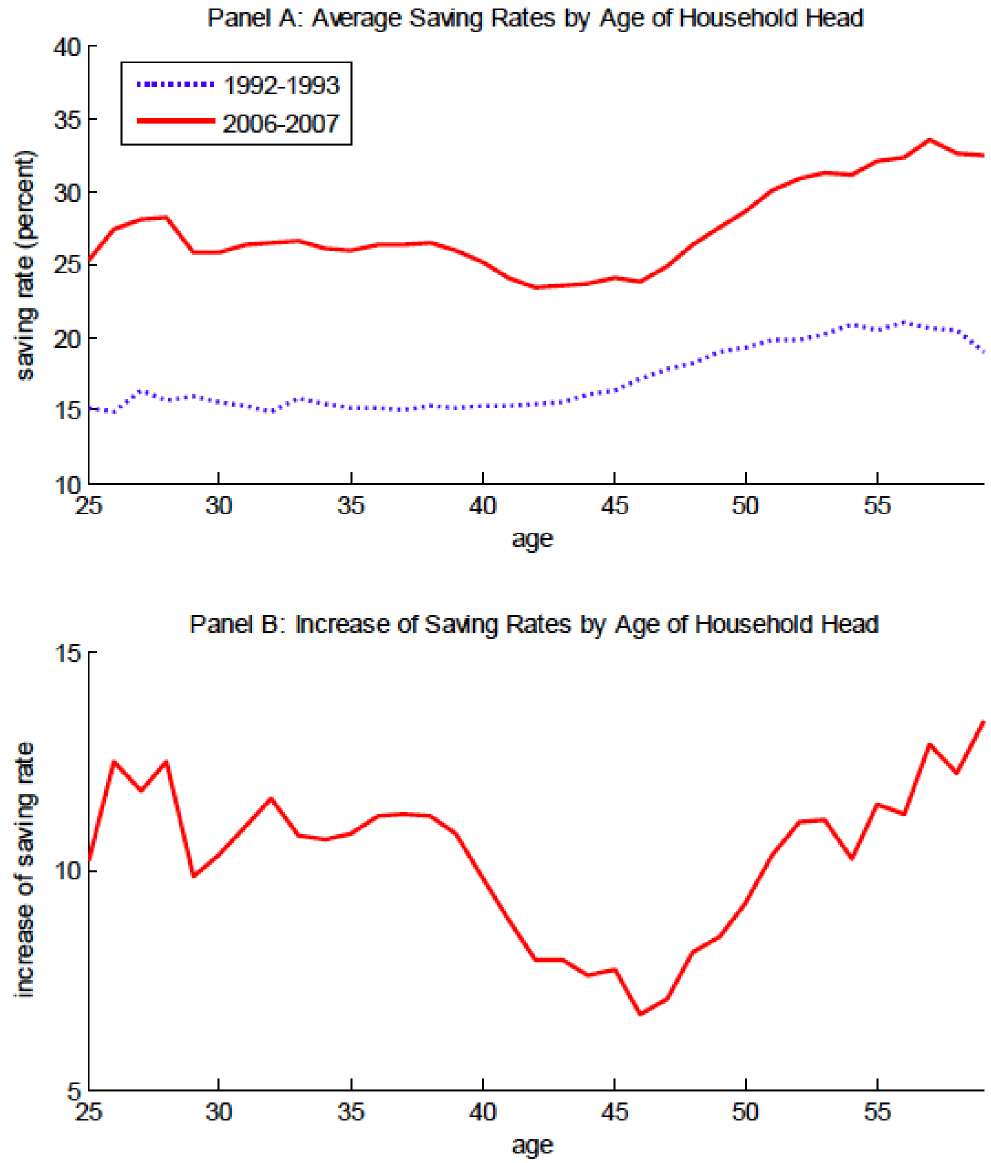

Figure 9: Cross-Sectional Age-Saving Profiles

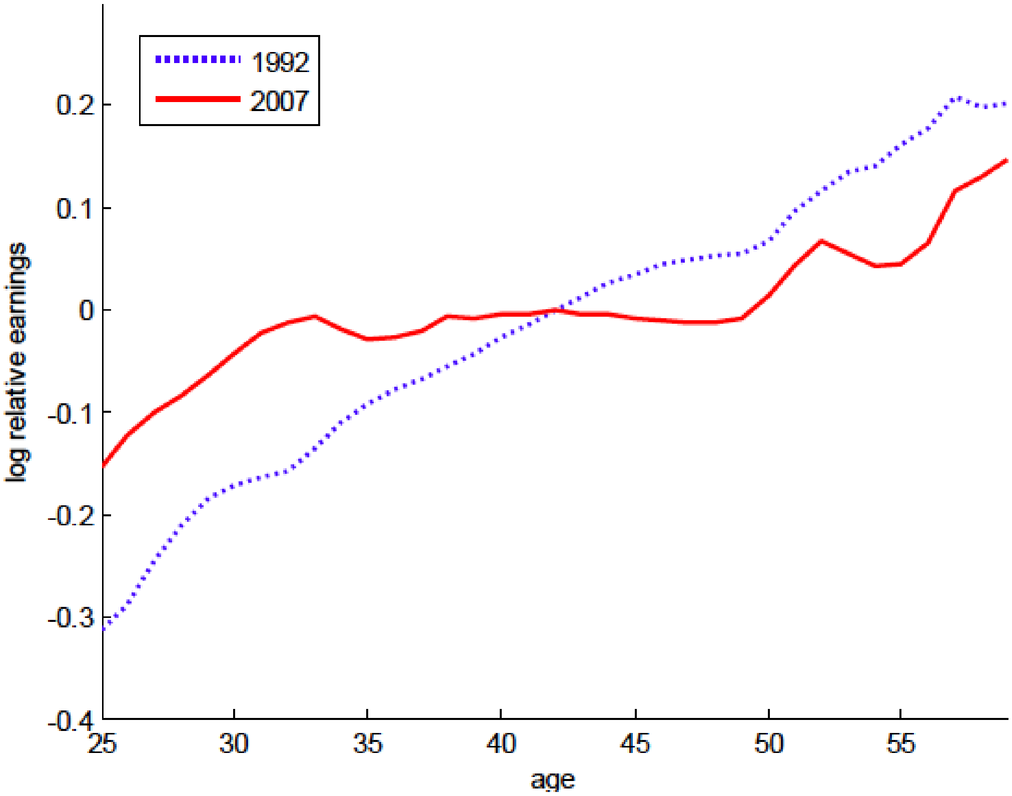

Figure 10: Cross-Sectional Life-Cycle Earnings Profiles
Relative earnings are computed as the ratio of earnings to earnings at age 42.

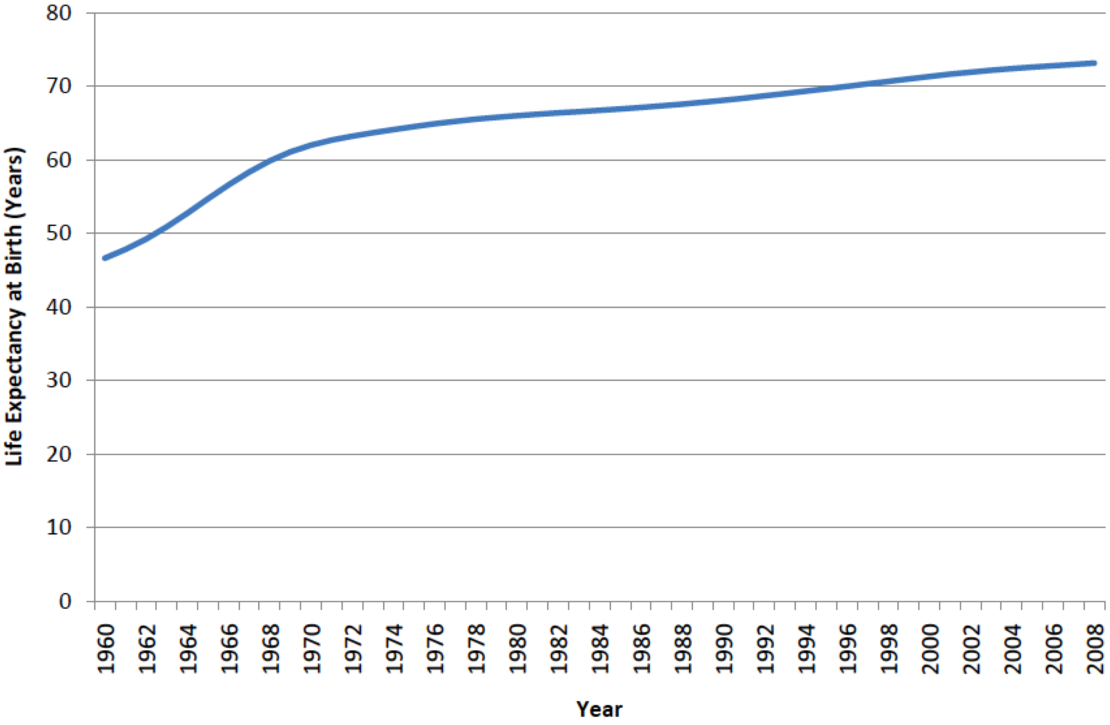

Figure 11: Chinese Overall Life Expectancy
Source: World Bank World Development Indicators.

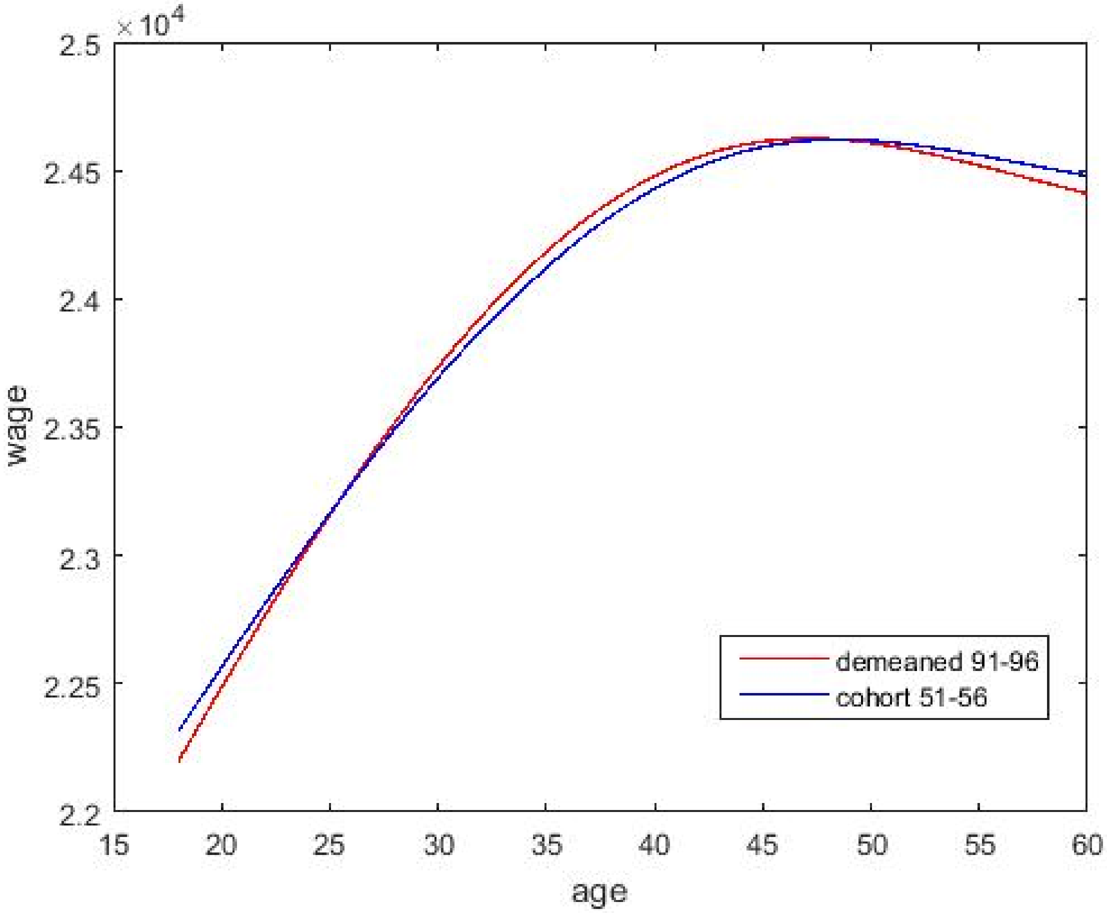

Figure 12: Simulated Age-earning Profile