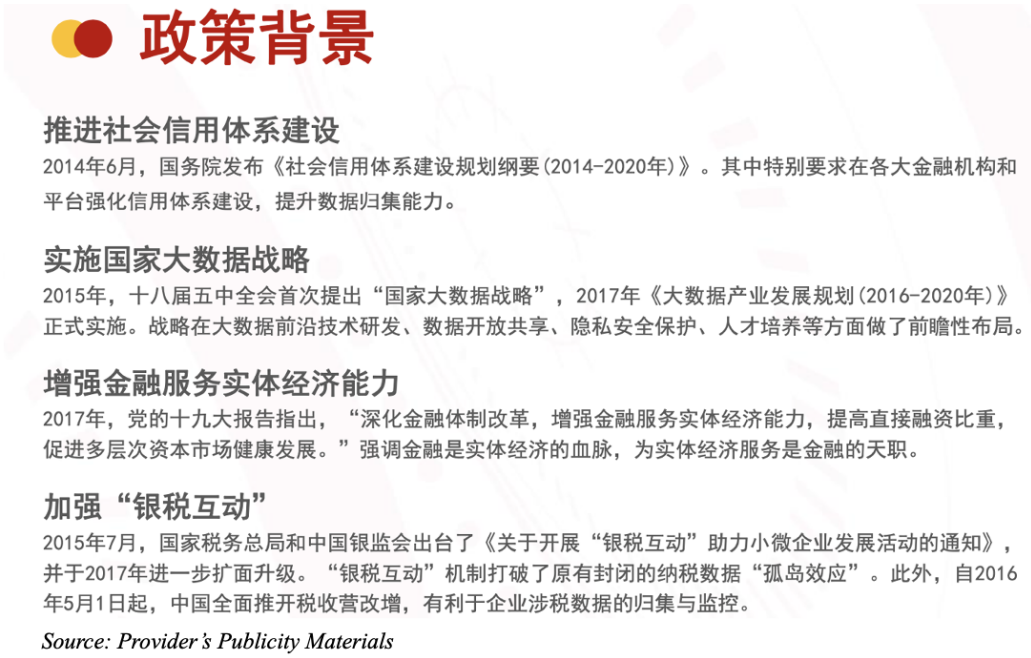
# Online Appendix

for "The Effects of Big Data on Commercial Banks" by Xiao Yin

## A. Policy Background




*Source: Provider's Publicity Materials*

### Policy Background

**Promoting the Development of a Social Credit System**
In June 2014, The State Council issued the *Outline of the Plan for the Construction of the Social Credit System (2014-2020)*. In particular, it requires major financial institutions and platforms to strengthen the construction of the credit system and improve the ability of data collection.

**Implementing the national Big Data strategy**
In 2015, the Fifth Plenary Session of the 18th CPC Central Committee put forward the "National Big Data Strategy" for the first time, and in 2017, the *Big Data Industry Development Plan (2016-2020)* was formally implemented. The strategy has made a forward-looking layout in the research and development of cutting-edge big data technologies, data open sharing, privacy security protection, personnel training and other aspects.

**Strengthening the Ability of the Financial Sector to Serve the Real Economy**
In 2017, the report of the 19th CPC National Congress pointed out that "we will deepen the reform of the financial system, enhance the financial sector's ability to serve the real economy, increase the proportion of direct financing, and promote the healthy development of the multi-tiered capital market." He stressed that finance is the blood of the real economy and serving the real economy is the primary duty of finance.

**Strengthening "Bank-Tax Interaction"**
In July 2015, the State Administration of Taxation and the China Banking Regulatory Commission issued the *Notice on Carrying out the Activity of "Bank-Tax Interaction"* to help the development of small and micro enterprises, which was further expanded and upgraded in 2017. "Bank-Tax Interaction" mechanism breaks the original closed tax data "autarky effect". In addition, since May 1, 2016, China has comprehensively promoted the replacement of business tax with value-added tax, which is conducive to the collection and monitoring of tax-related data of enterprises.

# B. Estimation Details

The structural estimation follows Crawford et al. (2018) and Ioannidou et al. (2022). There are two steps. The first step is price prediction, and the second step is the joint estimation of demand and default.

## B.1. Price Prediction

In the data, I only observe interest rates of the loans that are successfully originated. However, to estimate the indirect utility of demand, I need data on the interest rates of the loans from the banks that the borrowers do not borrow from. Therefore, the first step is a price prediction step. The target of this step is to get these unobserved interest rates. To do so, I adopt the following pricing model:

$$i_{j,k,t} = \gamma_0 + \gamma_1 d_{j,k,t} + \gamma_2 l_{j,k,t} + \lambda_{k,t} + \omega_i^q + \tau_{j,k,t},$$

where $d_{j,k,t}$ is the distance between firm $j$ and the nearest branch of bank $k$, $l_{j,k,t}$ is the log of loan amount, $m_{j,k,t}$ are dummies for maturity, where $\lambda_{k,t}+\omega_i^q$ are and bank-year and firm fixed effects, and $\tau_{j,k,t}$ are prediction errors.

To predict the interest rates offered to non-borrowing firms, I use propensity score matching on observable characteristics and then randomly assign a borrowing firm's fixed effect, $\omega_i^q$, to a matched non-borrowing firm. I assign the granted loan amount and maturity to non-borrowing firms using the same approach.

## B.2. Simulated Maximum Likelihood Estimation

I estimate the demand and default system using a two-step method based on maximum simulated likelihood and instrumental variables estimation. In the first stage, using data on firms' choices of bank and default, I estimate the firm-level and bank-level parameters across the two equations, $\eta = \{\alpha^D, \eta, \eta^D, \beta, \beta^D\}$, and the variance-covariance matrix of the errors in the system $\sigma$ and $\rho$. I also recover the bank-market-year specific constants (mean utilities) in the demand model ($\delta_{k,t} = \alpha_0 + \mathbf{X}_{k,t}\beta + \xi_{j,t}$) following Berry et al. (1995).

The indirect utility from demand can be written as

$$\begin{aligned} U_{j,k,t} = \delta_{k,t} &+ \alpha_i i_{j,k,t} + +\alpha_{i,Z} i_{j,k,t} \times Z_{j,k,t} \\ &+ \underbrace{\alpha_T T_{j,k,t} + \alpha_Z Z_{j,k,t} + \alpha_{T,Z} T_{j,k,t} \times Z_{j,k,t} + \mathbf{Y}_{j,k,t}\eta}_{V_{j,k,t}} + \epsilon_j + \nu_{j,k,t}, \end{aligned}$$

where $\nu_{j,k,t}$ is assumed to follow a T1EV distribution. Then the probability that borrower $j$ in year $t$ chooses bank $k$ is given by

$$q_{j,k,t} = \int \frac{\exp\{\hat{\delta}_{k,t} + \alpha_i i_{j,k,t} + \alpha_{i,Z} i_{j,k,t} Z_{j,k,t} + V_{j,k,t}\}}{1 + \sum_l \exp\{\hat{\delta}_{l,t} + \alpha_i i_{j,l,t} + \alpha_{i,Z} i_{j,l,t} Z_{j,l,t} + V_{j,l,t}\}} f(\epsilon_j) d\epsilon_j.$$

The probability of default conditional on borrowing is

$$p_{j,k,t} = \int \Phi_{\epsilon_i^D|\epsilon_i} \left( \frac{\mathbf{z_p} - \tilde{\mu}_{\epsilon_i^D|\epsilon_i}}{\tilde{\sigma}_{\epsilon_i^D|\epsilon_i}} \right) f(\epsilon_i|D=1) d\epsilon_i,$$

$$\begin{aligned} \mathbf{z_p} = \alpha_0^D + \mathbf{X}_{k,t}\beta^D + \alpha_i^D i_{j,k,t} + \alpha_T^D T_{j,k,t} + \alpha_Z^D Z_{j,k,t} \\ + \alpha_{i,Z}^D i_{j,k,t} \times Z_{j,k,t} + \alpha_{T,Z}^D T_{j,k,t} \times Z_{j,k,t} + \mathbf{Y}_{j,k,t}\eta^D, \end{aligned}$$

where

$$\epsilon_i^D|\epsilon_i \sim N(\underbrace{\rho\epsilon_i\sigma_D/\sigma}_{\tilde{\mu}_{\epsilon_i^D|\epsilon_i}}, \underbrace{\sigma_D^2(1-\rho^2)}_{\tilde{\sigma}_{\epsilon_i^D|\epsilon_i}}).$$

The two probabilities give the joint likelihood

$$\ln L = \sum_i \mathbf{q}_{j,k,t} \left[ \ln(\overline{q}_{j,k,t}) + \mathbf{p}_{j,k,t} \ln(\overline{q}_{j,k,t}^D) + (1 - \mathbf{p}_{j,k,t}) \ln(1 - \overline{q}_{j,k,t}^D) \right],$$

where $\mathbf{q}_{j,k,t} = 1$ if $j$ chooses $k$ at $t$, and $\mathbf{p}_{j,k,t} = 1$ if firm $j$ defaults.

## B.3. Identification

The identification strategy is based on the control function approach following Train (2009). Using this strategy is motivated by the fact that both demand and default are nonlinear models. This strategy consists of two steps. In the first stage, I regress the predicted and actual interest rates on the observables that are included in the demand and default models, plus a set of instrumental variables. In the second stage, I include the residuals from the pricing regression as control variables in the demand and default models to control for any unobserved factors correlated with prices. Doing so allows the identifying variation left over in prices to be orthogonal to demand and default unobservables[11]. Following Crawford et al. (2018) and Ioannidou et al. (2022), I use interest rates on households' deposits and the number of deposit accounts as the instruments. These variables serve as proxies for banks' funding costs.

[11] See Train (2009) and Wooldridge (2015) for more details.

# C. More Results

Table C.1: Risk Score and Screening Performance

This table gives the predictive performance of banks' proprietary risk score (Score) separately for the control and treatment groups and before and after the experiment. Risk score is standardized by each bank. The analysis focuses on the borrowers that have borrowed from both before and the experiment and both from a control bank and from a treated bank. The parentheses in columns (1) to (4) contain the standard errors. The $p$-value of the DID estimates in panel A is based on 500 Bootstrapping draws. The DID estimate in Panel B gives the difference-in-difference estimates between the changes in the AUC of the treated group and that of the control group, for which the $p$-value is calculated based on DeLong ER (1988). The sampling period is from 2015Q2 to 2019Q2. All variables are winsorized at 1% level by year-quarter.

| | Control | | Treatment | | |
|---|---|---|---|---|---|
| | (1) Before | (2) After | (3) Before | (4) After | (5) DID |
| | Panel B1: Logistic Regression | | | | |
| Score | 1.10 | 1.04 | 1.11 | 1.19 | |
| | (0.01) | (0.01) | (0.01) | (0.01) | |
| Pseudo R2 | 13.11% | 11.07% | 13.42% | 21.33% | 9.95% |
| | | | | | (0.00) |
| | Panel B2: ROC | | | | |
| AUC | 0.7643 | 0.7339 | 0.7622 | 0.8276 | 0.0958 |
| | (0.0076) | (0.0064) | (0.0077) | (0.0053) | (0.00) |
| N | 70,043 | 73,218 | 37,970 | 41,112 | |

## Table C.2: The Effects of the Policy by IT Intensity

This table gives the heterogeneous treatment effects of the policy on the loan-level variables by bank IT Spending before the experiment. IT-spending is banks' average IT spending to total expenses before the experiment. log Volume is the log of the amount of each loan in 10-thousands CNY. log Time is the log loan origination time in days. Interest is the interest rate (%) of the loan. Default is an indicator that the loan is defaulted. Regressions are weighted by loan volume. The sampling period is from 2015Q2 to 2019Q2. All variables are winsorized at 1% level by year-quarter.

| | (1) log Volume | (2) Interest | (3) log Time | (4) Default |
|---|---|---|---|---|
| Treat | 0.01 | 0.24* | -0.14* | -0.14 |
| | (0.02) | (0.13) | (0.08) | (0.12) |
| Treat × Hight IT | 0.01 | 0.57*** | -0.33*** | -0.37*** |
| | (0.02) | (0.11) | (0.10) | (0.13) |
| Observations | 222,343 | 222,343 | 222,343 | 222,343 |
| R-squared | 0.067 | 0.112 | 0.083 | 0.076 |
| Year-Qtr FE | Yes | Yes | Yes | Yes |
| Bank FE | Yes | Yes | Yes | Yes |

Standard Errors Clustered at Year-Quarter and Bank Level in Parentheses
*** $p$ <0.01, ** $p$ <0.05, * $p$ <0.10